VOLUME 61 WHOLE NO. 285
NUMBER 4 1947

Psychological Monographs

JOHN F. DASHIELL, *Editor*

# A Systematic Approach to the Construction and Evaluation of Tests of Ability

*By*

JANE LOEVINGER

*Washington University*

This monograph is essentially the same as a dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in the Graduate Division of the University of California, June, 1944.

*Published by*

THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

*Publications Office*

1515 MASSACHUSETTS AVE., N.W., WASHINGTON 5, D.C.

## ACKNOWLEDGMENTS

I WISH to thank the members of my thesis committee, Professors Warner Brown, Jerzy Neyman, and Frank N. Freeman, for encouragement and helpful criticism. I am indebted also to Professor Harold H. Carter for reading critically the entire manuscript. In particular, Professor Neyman suggested the formulation of the concept of homogeneity in terms of probability. A course in mental measurement under Professor Florence L. Goodenough, some years before the initiation of this essay, undoubtedly influenced my thinking in the direction here taken.

J. L.

# TABLE OF CONTENTS

# INTRODUCTION: THE PRESENT STATUS OF INTELLIGENCE TESTING

R. B. CATTELL (4) recently has summarized the present status of adult intelligence testing. He lists 44 adult intelligence tests, presumably all or nearly all written since the first World War; the list is impressively long, and it has been assumed widely that great advances in the theory and practice of intelligence testing have taken place since the first World War. And yet, Cattell continues, in a practical situation the number of available tests is always very small, and indeed, in the recent military situation it was found necessary to compile new tests or new batteries of tests. The psychologists called on to help in the selective programs in the second World War were in a position astonishingly like that of the psychologists who first drew up the Army Alpha; they were still improvising. They had the advantage only in more types of tests to choose from and in more data, of precarious relevance, on the various types. Cattell attributes the inadequacies of the present practices in adult mental measurement to the absence of fundamental discussions as to the nature of intelligence and the widespread "assumption that this field of research is a closed book."

Turning to the definitions of intelligence defended by or implicit in the papers of the authors of intelligence tests, Cattell finds the same definitions which were being proposed and severely critisized twenty and thirty years ago. The chief innovation has been the notion that by supposing intelligence to be composed of several special abilities rather than one general ability, the problem of defining intelligence is avoided; but the result of this assumption has been rather to reproduce the general confusion on a larger scale. "In short, a survey of current literature reveals every possible variety of divergence as to the objectives of intelligence testing. Intelligence is abstract thinking; it is concrete thinking; it is verbal skill; it is manipulative ability; it is innate; it is a set of acquired skills; it occurs equally in all activities; it cannot be measured by sampling; it is one thing; it is a host of things; it is a few distinct, clear-cut aptitudes."

With regard to the methods actually used in determining the validity of the tests, the situation is perhaps even worse. These methods are either tautological or more or less irrelevant. Validation against previous tests or estimates of intelligence or against the total test is tautological. The more or less irrelevant methods require the valid test to predict differences which are only partly intellectual differences, such as the differences between mental defectives and superior deviates, or differences in scholastic or occupational achievement; or they require the valid test to conform to some criterion which is also satisfied by completely nonintellectual characteristics, such as increase of score with age, or normality of score distribution. Of these criteria, Cattell finds most respectable "the uncompromising statement of a certain class of applied psychologist that he is not at all interested in intelligence but only concerned to establish a correlation between a test and certain kinds of life success." Besides the objection that intelligence and even special abilities are only partially responsible for success in life situations, there is the objection that, for example, different abilities may be called on for success in different schools, and technological changes in

a profession such as engineering may change substantially the abilities called on for success in that occupation. "Tests which need to be scrapped with every slight spatial or temporal change are not very economical. On the other hand, if it is only necessary to re-evaluate from time to time the 'real life' success situations in terms of a few unchanging reference factors, *e.g.*, intelligence, mechanical aptitude, verbal ability, for which there are standard tests of known validity, psychometrics has some claim to be a science."

"When one considers that a way out of this bancruptcy was indicated at least a generation ago, . . . one is amazed that any competent psychologist is content to continue discussing and 'investigating' tests in the limited language of clinical impression or within the shackling superstitions of educational tradition. Factor analysis does not bring in sight the end of all disputation, but it at least transports investigation to an objectivity far above that amateurish level of evaluating abilities which the psychologist has all too long been content to share with the layman." The apparent confusion within the field of factor analysis has been reduced by recent demonstrations of the equivalence of different factorial solutions, and the only major outstanding problem is the choice between the Thurstone method, emphasizing group factors, and the Spearman method, emphasizing a general factor. The solution proposed by Cattell, settling this difference by ballot, is by no means the only practicable one. A situation in some ways preferable, and certainly easier to enforce, would be that two schools of psychometry should develop, based on the two possible sets of assumptions concerning the nature of mental organization. It seems likely that before very long sufficient evidence on the relative efficacy of the different sets of assumptions should be accumulated to permit a rational and persuasive choice between them.

Cattell is not the only psychologist today calling attention to the present problems in the field of intelligence testing and suggesting a solution involving the application of factor analysis. As a single example, we may take the concluding statement of Guilford's (13) presidential address to the Midwestern Psychological Association in 1940: "If we are to continue to make effective progress in the discovery of new fundamental abilities and in their measurement, we should look to a more analytical study of individuals and of tests. And this implies not only factor analysis but also experimental analysis in combination with it. We need new tests of novel kinds, but we also need keen observations of human mental processes, particularly in the sphere of the symbolic activities."

The remainder of Cattell's article is devoted to discussion of other problems in the measurement of intelligence: the problem of the reliability or consistency of a test, the choice of subtests, the units in which the score should be expressed, and certain problems specific to adult testing, such as sampling, decline of abilities with age, and motivation. This thesis is concerned with the first three of these problems, the consistency of a test, the choice of subtests or items, and the unit of measurement. A comparison of the contents of Part I of this thesis with Gulliksen's (14) outline for a course in mental testing, will verify further that these three topics comprise a major and central portion of the theoretical aspects of test construction. The writer believes that precisely these three problems, besides being the major problems in test

construction *per se*, are *first* problems in the sense that their solution is presupposed by the most powerful instrument of test analysis now available, namely, factor analysis.

The remark must be appended to Cattell's discussion that to date the results of constructing tests on the basis of factor analysis have been disappointing. The tests so constructed have by no means been convincingly superior to all previous tests of abilities. Thurstone (35) has attributed his disappointment with many factorial studies to the failure of the studies to live up to certain fairly technical requirements of factor analysis. Thurstone's works are, of course, not subject to the criticism of failure to observe the technical requirements of factor analysis; yet the issuance of his Primary Abilities Test provoked almost as much disappointment as praise (3). Two psychologists suggested that the source of the difficulties may lie partly in the original tests. Kelley (3) said, "It is necessary that pragmatic values and not wishful thinking—even though introduced in so inconspicuous a manner as in the selection of the original tests to be subjected to factor analysis—be the basis of trait analysis." (p. 259)

A more direct emphasis on the importance of the original tests was evident in the view of Stalnaker (3) that, "At the moment . . . the greatest need seems to be for vastly improved basic tests, tests developed with but one object in view—a thorough and dependable measure of certain types of ability. Such tests will require more time from the candidates; some of the tests may be more difficult to score; all of them will require great care in construction. Until such tests are used as a source for the primary data, positive conclusions from any system of factor analysis must be viewed as conjecture." (p. 261)

To implement this last statement of Stalnaker's, consider the fundamental assumption of factor analysis. This assumption is that the score of each individual on each test is the weighted sum of his standard scores on each of the factors entering into performance on the test. The factor scores are assumed to be characteristics of the individual, constant for all tests. The weights in the summation are assumed to be functions of the test, constant for all individuals. People who have different scores must have done different items correctly. Even people who have the same scores often will have done different items. The assumption of constant weights imply that the various items of the test are measuring more or less the same complex of abilities. While allowance is made in factor analysis for a degree of unreliability of tests, no allowance is made for inconsistency in the sense, for example, that people with high scores are differentiated mainly with respect to verbal facility, while people with low scores are differentiated mainly with respect to verbal reasoning. Consistency with respect to ability measured, while related to reliability, is not measured adequately by traditional measures of reliability nor by recently proposed substitutes for the reliability coefficient. Rigorous application of the rule "consistency with respect to ability measured" will also help to clear up the chaos of rules and criteria for item selection.

Factor analysis, furthermore, depends on the correlation between tests to discover the relations of the abilities underlying the tests. It will be shown in Chapter VI, Section A, that two tests of the same complex of abilities, or even two tests of a single ability, can have any correlation greater than zero, depending

on the choice of the unit of measurement. The correlation between two tests of different abilities, can, in the extreme, be changed from positive to negative by altering the unit of measurement. Again it follows that a large element of confusion is introduced into factor analysis by ignoring the problem of the unit of measurement in the tests used.

The contention has been made here that the present difficulties in the measurement of mental abilities go deeper even than the failure to obtain adequate definitions of these abilities. Factor analysis offers a theoretical solution to the problem of defining the abilities to be measured, but factor analysis presupposes the solution of certain other problems of test construction. The accepted solutions to the problems of reliability, selection of items, and the unit of measurement will be shown each to depend on a new set of assumptions, and each to depend on assumptions which are remote from if not contradictory to the realities of the clinical situation in testing. An alternative set of solutions will be outlined, calculated to provide tests suitable for factor analysis, and based on a minimum number of assumptions, closely representative of the real situation in testing and common to all of the solutions. While the same problems may arise in connection with various types of psychological measurements, the present discussion will concern only tests of mental ability, as opposed to achievement and personality tests.

# PART I: SOME PROBLEMS IN TEST CONSTRUCTION AND EVALUATION

## CHAPTER I

## THE PROBLEM OF RELIABILITY

*A. The measurement of reliability.*

RELIABILITY refers to the self-consistency or self-correlation of the test, the extent to which it measures what it measures rather than transitory and irrelevant factors. The self-correlation of the test cannot be determined directly, but is estimated in one of three ways: the test-retest method involves administering a single test twice to the same group and computing the correlation between the two sets of scores; the split-half method involves correlating scores on half the items in the test against scores on the other half of the items; the comparable forms method involves correlating scores on one form of a test with scores on a second closely comparable test.

Assuming that the test is administered and scored without error, the sources of unreliability are of two types, accidental content factors and transitory variations in efficiency. Transitory variations in efficiency come from such factors as motivation, momentary set, fatigue, boredom, and illness. To the extent that these variations in efficiency are different for different items, they are inseparable from accidental content factors. As an example of error due to accidental content, we may consider a test containing a number of problems in verbal deductive reasoning. The problems would differ in subject matter, and apart from the difficulty of the relationships involved, some topics would be easier for a particular individual, relative to other individuals, than other topics, whether because of familiarity, congeniality, or specific emotional factors. Or we may think of accidental content factors in another way, namely, that some items will involve additional or slightly different abilities than other items or the test as a whole.

It is the function of the reliability coefficient to estimate the extent of the influence of these sources of unreliability. The greater the proportion of test variance caused by transitory variations in efficiency and by accidental content factors, the lower should be the reliability coefficient. The correlation of the score on odd items and the score on even items, which is the most usual way of computing the split-half reliability, is lowered by accidental content factors. As Goodenough (10) has shown, however, it is actually raised by transitory variations in efficiency, since the momentary level of efficiency enters both scores of each individual in the same sense. Similarly, transitory variations in efficiency enter the test-retest reliability coefficient in the appropriate sense, that is, unreliability due to transitory variations in efficiency lowers the test-retest coefficient. The effects of accidental content factors, however, are repeated in the retest score, and thus must act to raise rather than appropriately to lower the test-retest coefficient. If the comparable forms of a test are too closely similiar, the same criticism applies to a lesser degree. It appears that to have both sources of unreliability affect the reliability coefficient in the appropriate direction, one must administer on different occasions com-

parable forms of a test, constructed so that they measure the same ability but do not repeat accidental content factors. No well-known discussion of reliability has yet explained how one knows when he has constructed such a pair of tests.

When reliability is estimated by correlating two tests administered on different occasions, whether they are the same or comparable forms, there may be additional systematic factors influencing the reliability coefficient but not truly influencing reliability. If the retest is with the same form, the subjects will remember some of their original answers, and if the forms are closely comparable, they will at least have learned methods of solving the problems, in many cases. Besides direct learning and memory, which may act either to raise or to lower the reliability coefficient, there are such factors as interpolated practice on similiar materials, and if the time interval is long enough, actual changes in the abilities measured. In general these influences operate to raise the average level of performance on the second test, but their influence on the reliability coefficient will vary with special circumstances of different tests. When reliability is estimated by using two tests or two parts of a test administered on a single occasion, there is the possibility of fatigue and boredom affecting the second half of the test period more than the first half. For this reason the split-half reliability is seldom computed by correlating the first half of the test against the last. In the case of the split-half coefficient, moreover, there are two additional difficulties. The reliability of half the test is not the same as the reliability of the whole test; so the split-half coefficient requires a correction derived from the statistical theory of reliability. There are as many split-half correlation coefficients as there are ways of dividing the test in two; this ambiguity has led to the search for unique equivalents of the split-half correlation.

It is plain, at any rate, that the meaning of the reliability coefficient varies markedly with the manner in which it is determined, and that there is no entirely satisfactory way of determining the self-correlation of a test. But if there is no correlation coefficient which corresponds to it, what exactly is meant by the self-correlation of a test? It is here that we encounter the statistical theory of reliability.

### *B. The statistical theory of reliability.*

The basic assumption of the statistical theory of reliability is that the obtained score of an individual on a test can be expressed as the sum of his "true score" on the function tested and a "chance error" component. The reliability coefficient as ordinarily computed is obviously the correlation between two sets of obtained scores. No matter how the reliability coefficient is computed, the statistics of reliability assume that (a) the variable error factor has an expected or average value of zero, (b) the error factor in one set of obtained scores is uncorrelated with that in another set, however similar the tests may be, (c) the error factor in a set of scores is uncorrelated with the true scores, and (d) the variances of the error factors in two comparable tests are equal. On the basis of these assumptions, the reliability coefficient is shown to be equal to the ratio of the variance of true scores to the variance of obtained scores. In terms of the statistics of reliability, this is what the reliability coefficient means. The same assumptions are made in deriving the formulas for the standard error of measurement, the Spearman-Brown formula for estimating the reliability of a lengthened test, the

correction for attenuation, and various other formulas. The basic equation of the theory of reliability is, then, that the obtained score is the sum of the true score and a chance factor; and the true score must be identified in such a way that the difference between it and the obtained score will have the above four properties.

Thurstone (34) has written, "The true score in the test is assumed to be the average score that a subject would make in an infinite number of parallel forms of the test. Of course, the true score can never be actually obtained because the number of parallel forms that can be given to a person is finite and hence there will always be a residual of chance error even if we ignore the large systematic errors of fatigue and boredom which an attempt would necessarity invite. But theoretically the concept of a true measure as the mean of an infinite number of repeated measurements is a very useful one. Evidently, when a test is given to a subject we want to ascertain as nearly as possible his true score."

Concerning the definition of the true score as an average, the fact that only a finite average can ever be computed does not in any way invalidate the definition in terms of an infinite average, so long as it is clear that the "residual of chance error" approaches zero as the number of tests increases. Systematic errors such as fatigue, boredom, practice, and learning are a more serious objection to this definition. If we could be sure that the effects of these factors were systematic in the sense of being constant errors for all levels of ability or even for all persons at any level, it would be simple to correct the scores. One would need merely to subtract from each score the average gain of the group. But it is hardly conceivable that all persons would be affected the same way by boredom or by the opportunity to learn from taking a test or even by the whole complex of factors that operates in repeated measurings. The objection, be it noted, is not that the true score cannot be computed, but that it does not exist; it is defined in terms of operations, but operations which in the nature of things cannot be performed, namely, the averaging of repeated tests where there is no effect of repetition.

There is an even more serious objection to the definition of the true score as an average than the fact that it cannot be applied to obtain even a finite approximation: The definition of parallel forms is bound to bring us around to a full circle. Let us summarize the argument so far. There are several ways of computing a coefficient of reliability, but all of them yield only estimates of the true reliability or self-correlation of a test. We must therefore mean something different by reliability than any of these correlation coefficients. Expressing each test score as the sum of the true score and a chance error, it can be shown that the reliability coefficient which we are trying to estimate is the ratio of the variance of true scores to the variance of obtained scores. The true score is the average score of the individual on an infinite number of parallel forms of the test. Parallel forms of a test are most commonly defined as tests on which the true scores are equal. This definition is unsatisfactory because it is obviously circular with the definition of true scores. Kuder and Richardson (20) have defined parallel forms as tests composed of items paired so that the two members of a pair are equal in difficulty and correlated to the extent of their respective reliabilities. Thus test reliability is defined in terms of true scores, true scores in terms of parallel forms, and parallel forms in

terms of item reliability. The only way out of the circle appears to be Kelley's (18) definition of parallel forms in terms of the judgment of the test constructor that the tests are equally excellent measures of the same function or functions. Kelley then proceeds to assume that this judgment can be made perfectly successfully, but the whole weight of evidence from correlations and factor analyses is that the investigators cannot prejudge successfully the similarity of functions measured by tests or items. The definition of parallel forms is of less consequence than the manner in which they are constructed. In constructing two tests of the same mental function, the psychologist does perform the judgment which Kelley describes, namely, selecting items which look as if they would require the same mental abilities. But no psychologist would consider the appearance of similarity as sufficient to establish the tests as parallel forms; one would have to establish a high correlation between scores on the two tests as final evidence. We have defined reliability in terms of true scores, true scores in terms of parallel forms, and parallel forms in terms of correlation. Since the correlation between parallel forms is exactly the "reliability" whose meaning we are seeking, the definitions are again circular.

The true score on a test is not, however, always identified as an average. In the passage quoted above, Thurstone says that the true score is what we want to ascertain about the individual. Now in the case of testing abilities, a good many psychologists hold the view that what we want to ascertain is not what the individual usually does, but what he can do at his best. Much of the construction and most of the administration of intelligence tests is devoted to eliciting the optimal performance from each individual. It is understood, of course, that by optimal performance is meant the best performance consistent with the abilities of the individual and with the rules for administering and scoring the test. The notion of optimal performance is implicit in the idea of ability as contrasted with achievement or with personality traits. The discussion of the sources of unreliability agrees with the conception of the true score as an optimum; the difference between the optimal score and the obtained score is accounted for in terms of just the sort of factors which are included under the heading of sources of unreliability, that is, fluctuations in efficiency due to fatigue, boredom, illness, accidents of mental set, emotional blockages due to special content. The obtained score actually may be higher than the theoretical optimum, since it is sometimes possible for the individual to give the right answer for the wrong reason. In a well constructed test this possibility is usually not great. In general, the optimal score will be higher than the obtained. If we consider the optimal score as the true score, then the error factor has a negative expected value. The clinical conception of the true score as an optimum is consistent with the statistical conception of the true score as an average only if we assume that the difference between the optimal score and the average score is the same for all individuals tested. This difference, a constant error for all persons, would not affect the statistical derivations.

The error in score of which a clinical psychologist can obtain at least some intuitive, semi-empirical information is the difference between the optimal score and the obtained score. Probably few clinicians would maintain that the "clinical error" has the characteristics which the "statistical error" factor is supposed to

have, particularly the zero correlation for two closely similar forms and for error and true score. Nor would many clinicians agree that the difference between the "clinical true score" and the "statistical true score" is constant, which is the broadest assumption under which the clinical considerations can be consistent with the statistical derivations. Certainly there is no evidence to support this assumption, and such evidence would be exceedingly hard to produce in view of the fact that the statistical true score is defined in terms of operations which cannot be performed.

Surprisingly enough, few psychologists have questioned the validity of the assumptions underlying the theory of reliability; in recent years the tendency has been rather to invent new techniques based on the old assumptions plus additional ones. Around 1910, however, Brown carried on a discussion with Spearman on just these assumptions, in relation to the validity of the correction for attenuation. Brown and Thomson (2), repeating the criticisms at a later date, cited evidence from Pearson (23) showing that errors of measurement may very well be correlated on separate occasions. They point out that the "error factor" in mental measurement represents essentially individual variability rather than a true error of measurement, and that assuming the errors in one test uncorrelated with the errors in another test or with the mean value for the function is purely gratuitous. If the correction for attenuation is applied when the errors actually are correlated, they show by a hypothetical example, the obtained correlation is made more erroneous rather than less. "This undesirable phenomenon would seem to be particularly liable to occur when attenuation corrections are made by splitting a test, since a boy who is off colour in one part is very likely to be also off colour in the other and hence errors will be correlated." They then present evidence which they consider to be contrary to the assumptions in question for some simple types of judgments and problems.

Spearman's (28) reply to similiar criticisms had been, "Clearly, such assumptions are far from carrying conviction *a priori*." He admitted the possibility of systematic errors in addition to the type of "accidental" ones corrected by the formula. To eliminate systematic errors, other methods, such as better data or partial correlation, must be utilized. In order that the split-half reliability be used in correcting for attenuation, it must be determined from halves so made up that the differences in scores on the two halves can be considered "accidental." But Spearman offers no further practical advice on how to accomplish such a split.

Kelley (18) has recently espoused a point of view similar to that of Spearman. He states, "Unlike the correlation coefficient, which is merely an observed fact, the reliability coefficient has embodied in it a belief or point of view of the investigator," namely, that the two forms or two halves of a test correlated are "equally trustworthy measures" of the ability or abilities involved. Like Spearman, Kelley recommends that before computing a reliability coefficient, the investigator face squarely the act of judgment involved in declaring the two forms or two halves equivalent. And like Spearman, Kelley has nothing further to say in aid of those called on to make this judgment. He then goes on to derive the coefficient of reliability as an expression of the proportion of the variance of the differences between individuals which "is trustworthy or predictable from a knowl-

edge of the true differences," under the assumption that the investigator has been completely successful in constructing "equally excellent measures" of the function. Kelley points out that his derivation does not involve any assumptions about the correlation between errors, but at the beginning of his article he introduces the reliability coefficient in terms of properties which it possesses only under the assumptions of the statistical theory of reliability.

One criticism applies to the recommendation of intelligent judgment, both in the case of Spearman and that of Kelley. The investigator has no direct, verifiable knowledge of the factors concerning which he is to judge. How can a test constructor divide a test so that the errors in either half will be accidental, when he knows only the obtained score and has seen the errors only through the dark glass of the reliability coefficient? The task of dividing a test into two equally excellent measures is surely no different.

The conclusion is forced that the concept of reliability as at present defined and used is highly unsatisfactory. When two tests are used as equivalent forms, that is, when a score on one test is used as if it means the same thing as the same relative score on the other test, certainly the correlation between the two tests is a valuable piece of information. This correlation is not in any way clarified by the term "reliability". The test-retest correlation similarly will give valuable information in certain types of situations, but this coefficient should never be confused with the parallel forms coefficient by use of the common name of reliability. The split-half reliability measures again something different, and attempts to find a unique coefficient describing essentially the same thing as is described by the split-half correlation will be described in the next section. The statistical formulas utilizing reliability coefficients are based on assumptions at best so inaccessible, at worst so contrary to clinical experience, that the attempt to find a substitute for the notion of reliability, based on assumptions closer to the real situation in testing, appears well justified.

Radical as this conclusion may seem, it is not original. On the basis of considerations some of which have been quoted above, and some additional empirical considerations, Goodenough (10) concluded: "What we should do, I think, is to relegate the use of the term 'reliability' to the limbo of outworn concepts and express our results in terms of the actual procedure used. It is quite as easy to speak of the 'correlation between test and retest' after a stated interval or between 'the sums of alternate items' or between 'equivalent forms' of a test as it is to use the more conventional but far less accurate expression, 'reliability.' . . . By giving up the name while retaining the processes we shall gain in precision of thought and expression with no loss of informative data." It is deplorable that Goodenough's clearly reasoned article has been so seldom referred to during the decade following its publication, during which time the concept of reliability has received so much algebraic elaboration.

### *C. The "rational equivalence" method of determining reliability.*

Recently much attention has been focussed on the problem of finding a unique estimate of reliability on the basis of a single administration of a test, without requiring any such judgment as was called for by Spearman or Kelley. One formula has been arrived at by several investigators using different methods;

and its wide acceptance has been further accelerated by the fact that it is much easier to compute than a split-half reliability coefficient.

Kuder and Richardson (20) derived several approximations to the reliability coefficient, all of which are based on the definition of equivalent forms in terms of the interchangeability of the items in pairs: "The members of each pair have the same difficulty and are correlated to the extent of their respective reliabilities. The inter-item correlations of one test are the same as those in the other." On the basis of additional assumptions they derived formulas for the correlation between equivalent forms in terms of the intercorrelations of the items and in terms of the correlations of the items with the total score. These formulas have not come into wide use, probably partly because both require more work than the split-half reliability. Their most widely quoted formula is their formula (20) which requires only the variance of the test ($\sigma_t^2$), the number of items ($n$), and the average variance of the items ($\bar{pq}$).

$$r_{tt}=\frac{n}{n-1}\cdot\frac{\sigma_t^2-n\,\bar{pq}}{\sigma_t^2}$$

They derived this formula on the explicit assumption "that the matrix of inter-item correlations has a rank of one and that all intercorrelations are equal," apparently under the impression that this assumption is consistent with allowing item "difficulties to vary over a wide range." Jackson and Ferguson (17) have pointed out that their explicit assumption implies that all items are equal in difficulty. Midway in Kuder and Richardson's derivation, they make the further assumption of "equal standard deviations of items," which is equivalent to the assumption that there are at most two degrees of difficulty of items, that is, the number passing any item must equal either the number passing or the number failing any other item.

Dressel (5) has shown that, apart from its derivation as an approximation to a reliability coefficient, Kuder and Richardson's formula (20) "measures the homogeneity of the items in a test, having the value 1 if the items are perfectly intercorrelated with equal variances and the value 0 if the items are mutually independent or if a number of the items are negatively discriminating." While he does not prove that the formula will yield a value of one *only* if the items are perfectly intercorrelated and have equal variances, this statement could be proved easily from the form of the equation which he presents. The correlation between two items is not an entirely unambiguous idea, but looking at the derivations, clearly Dressel and Kuder and Richardson imply the "four-point correlation," which is computed by straightforward application of the product-moment coefficient to a four-fold table. And the observation of Jackson and Ferguson still holds true, that the correlations can equal one only between items equal in difficulty. From the statement that the reliability will equal one only if all the items are perfectly correlated and equal in difficulty, it is only one step to the statement that the reliability will equal one only if everyone has a score of zero or perfect. For everyone who passes one item passes all, and everyone who fails one fails all. Exactly as good results could be obtained by giving just one item instead of the whole test. Thus, according to this formula, a test can have a reliability of one only in an absurd instance.

The reason that the formula works in

this way is that it is derived from the reliability coefficient under the assumptions that items are of equal difficulty and intercorrelate to the extent of their reliabilities, that is, the correlations between items corrected for attenuation are assumed to be unity. Dressel's derivation shows that the reliability coefficient so derived is lowered by any inconsistency between these assumptions and the actual case, as well as by the unreliability of the items. Since we desire to have reliable tests, and since no one would use a test on which everyone scored either zero or perfect, we may conclude that the Kuder-Richardson formula applies only to a case of no importance. The same formula is derived by Hoyt (16) and by Jackson and Ferguson (17), on the basis of two new sets of assumptions, but since the result still has the same consequences, both derivations are suspect of harboring the original or equally bad-assumptions.

Hoyt derives the Kuder-Richardson formula from the definition of the reliability coefficient as the ratio of the variance of true scores to the variance of obtained scores, using analysis of variance and the method of Markoff's theorem. His initial assumption is that the error component for each person on each item is normally distributed with the same variance as the error component in every other item. The error component is defined as the difference between the actual score and the true score of the person on the item. The true score is a constant based on the difficulty of the item and the ability of the person. Since the actual score on the item is either one or zero, and the true score is a constant, the error component must equal either one minus the true score or simply minus the true score. The error component for any one person and any one item has only two possible values, which is a far departure from the normal curve. Moreover, the variance of the error component depends solely on the probability of the person passing the item; so the assumption of a constant variance for the error component is equivalent to the assumption that the probability of any person passing any item is a constant. Hoyt's assumptions are worse than Kuder and Richardson's; rather than simply restricting consideration to an unimportant special case, Hoyt has considered an impossible case, for his assumptions are mutually contradictory.

Jackson and Ferguson supply the most innocent-looking derivation. They derive formula (20) as an expression for the correlation between two equivalent tests using only the definition of equivalence as the property of tests equal with respect to average item variance and average inter-item convariance. At one step in their proof, they assume, without so stating, that the sum of two equivalent tests will have the same average inter-item covariance as the two tests separately. The implications of this assumption are not obvious, but they report the finding that the average inter-item covariance has a lower limit which is a function of the number of items, though the upper limit is independent of the number of items. One may suspect that in general adding two tests with same average inter-item convariance will *not* result in a test with the same value.

Surely one must be disturbed to find six psychologists offering four separate derivations of a formula as valueless as this one, not to mention the many laudatory articles about the formula written by others. Among the latter must be mentioned Froehlich's (9) article, which suggests the use of the formula by people without enough statistical training to compute a correlation coefficient!

The formula has also been criticized, notably by Kelley (18). In connection with Kuder and Richardson's formula (21), which differs from formula (20) only in carrying one step further the consequences of assuming items of identical difficulty, Kelley showed that adding a number of items which every one could do correctly would change markedly the reliability according to the formula, although not changing at all the reliability in the ordinary sense. Kelley does not explain why the formula behaves this way, but from the above considerations it is obvious that the reason is that the heterogeneity of difficulty introduced with the easy items is reflected as a decrease in reliability.

Wherry and Gaylord (37) have also criticized the various formulas of Kuder and Richardson. They feel that a more reasonable set of assumptions would be that each test is made up of groups of items such that within each group the item variances are equal and the inter-item correlations are all equal to the item reliabilities. They further assume that the factors measured by the groups will be mutually orthogonal. They admit in a footnote, "We deliberately chose to magnify the discrepancy between our results and those based upon the assumption of a single factor, due to our feeling that the actual usual case would lie somewhere in between the two extremes." Thereafter, however, they refer to the reliability formula based on their assumptions as the "true formula" and show that by comparison the Kuder-Richardson formulas are in error.

The alternative which Wherry and Gaylord offer to the method of rational equivalence is scarcely more acceptable in terms of the assumptions involved. They make the highly restrictive assumptions of equal difficulties and of correlations to the extent of reliabilities for the items within a subtest, plus the assumption that items in different subtests will have zero correlations with each other. Their coefficient is in consequence not a satisfactory answer to the need for a coefficient describing the internal consistency of a test. Kelley suggests a quite different alternative. He considers the measurement of the "unity or coherence of a test" to be an important problem somewhat different from the problem of reliability, and not satisfactorily answered by the Kuder-Richardson proposals. As a possible solution, he suggests a factor analysis of the items of the test, and the computation of a "coefficient of coherence" which is essentially the ratio of the variance of the first component of the factor analysis to the sum of the item variances. He admits, however, that the computations involved would be enormous if the number of items were at all large. In Chapter IV, Section B, a coefficient of the homogeneity of a test will be proposed which appears to describe exactly the characteristic of the test which Kelley described by his coefficient of coherence. The coefficient to be proposed here has two advantages over Kelley's. It is very easy to compute, involving ordinarily less work than a split-half correlation, and it is more closely related to the "singleness of purpose of the items constituting the test." Kelley's coefficient cannot be computed without first computing correlations between items, which will be shown in Chapter II, Section C, to involve a new set of difficulties, and then making a factor analysis, which introduces a number of additional assumptions.

## Chapter II

## The Problem of Item Selection

### *A. The principle of item selection.*

Every mental test in use is made up of separate items or subtests. The problem of how to select the items is one of the major problems of test construction, and astonishingly different principles have been propounded as the basis of item selection. Two of the most widely held will here be called the Sampling principle and the Equivalence principle. The Sampling principle states that the correlation of items with criterion should be as high as possible, and the correlation of items with each other should be as low as possible. This principle seems to be based on the sampling theory of testing intelligence. The Equivalence principle states that the correlations of the items with each other should be as high as possible and the items should be as nearly as possible of the same difficulty. Items so chosen will tend to be equivalent to each other.

Thomson (30) has made a similar distinction, in these terms: "Validity is concerned with whether a test measures the right thing, discrimination depends on whether the test spreads the candidates out well. When there is a recognized criterion or independent measure of what is tested, the validity of the test is estimated by its correlation with that criterion. In this case the mathematics of the regression equation tells us which items to add to improve the validity. They are those items . . . which correlate well with the criterion but badly with the battery. . . . But just because such an item correlates badly with the preexisting battery it will not be increasing very much the discriminating power of the battery among men." Many investigators tend "to forget the criterion and to add items which correlate very well with the existing battery, and increase its discriminating power, its natural standard deviation." Investigators who are concerned to increase the correlation of the test with a criterion are following what is here termed the Sampling principle. It will be shown shortly that disciples of the Equivalence principle will increase the standard deviations of their tests but at the expense of true discriminating power. An alternative principle, aimed at increasing the true discriminatory power of the test, will be developed in Chapter V, Section A. Thomson is mainly concerned with predicting a criterion, and for this case he offers the alternative principle that as new items, correlating closely with the criterion, are added to the test, items with poor criterion correlations should be dropped so as to leave items closely related to each other, thus not impairing the discriminating power of the test.

An idea of how widely the Sampling principle of item selection has been accepted can be obtained from the following passage from Greene (11): "Most of the early examiners, like those of today, followed the hypothesis that persons were possessed of a general faculty called intelligence, which could be measured by a variety of mental tests. They usually wished to appraise intelligence for practical predictions of some sort. They selected only those items for an intelligence test which showed fairly high correlations with some criterion of intelligence and low correlations with each other. An important application of this method of selection was the development of United

States Army mental tests in 1917." (p. 301) Results from correlating subtests on the Army Alpha, however "led empirically to the conclusion, which is also mathematically obtainable, that it is impossible to secure subtests which will correlate highly with a criterion, and nearly zero with each other." (p. 307) Cattell (4), who had the apparently erroneous impression from another publication that Greene approved the Sampling principle of item selection, commented, "Whatever else we may feel about this criterion, which is widely entertained, though less explicitly, by several clinical psychometrists, we have to add that its complete consummation is a logical and mathematical impossibility."

If some psychologists believe that selecting items according to the Sampling principle is undesirable or ultimately impossible, others of equal statistical sophistication accept the principle entirely. Flanagan (8) considers that the multiple regression formulation is a "precise mathematical solution" of the problem of item selection, and he believes that test statisticians in general will agree. As Thomson pointed out, the mathematics of the regression equation imply just the rule for item selection which we have called the Sampling principle. Flanagan seems to feel that the only difficulty which remains in the field of item selection is that when there are a large number of items, the amount of work by the multiple regression method is prohibitive.

Horst (15) has recently worked out a method of item selection based on the principle of maximizing the correlations of the items with a criterion. In order to arrive at even an approximate solution, Horst makes an assumption which he admits is not necessarily true, but which he does not state in a sufficiently accessible form so that an impatient reader can make any estimate of its reasonableness. The assumption appears to be approximately the statement that those items which have at once a maximum correlation with the criterion and a minimum correlation with each other will be the same items which have at once a maximum correlation with the criterion and a minimum correlation with the original test. With more assumptions and approximations, a procedure is then evolved which depends on selection of items with low correlations with the original test to provide items with low correlations with each other. Since these items are also selected for having high correlations with the criterion, they will probably have something in common with each other that they do not have in common with the rejected items of the original test; in that case the assumption will be wrong and the rationale of the method will break down. Even allowing the acceptibility of this assumption, the method is tedious and approximate.

One criticism and one restriction to the Sampling principle may be added. The criticism is that the use of the multiple regression equation for item analysis assumes that the correlation of an item with a test is a perfectly clear idea. The theory of correlation has been developed mainly for expressing the relationship of two or more many-valued variables, while an item score has just two possible values, zero or one, in the cases discussed above. The biserial coefficient of correlation is the equivalent of the Pearson correlation on the assumption that the dichotomously scored variable is actually continuous, if certain other conditions hold true. The use of this coefficient introduces a number of gratuitous assumptions. The coefficient which is probably implied in the use of the multi-

ple regression method of item selection has been called "point biserial r", and is computed by a straightforward application of the Pearson product-moment formula, without regard to the restriction concerning many-valued variables. Point biserial r does not even have the elementary virtue of having a maximum value of one, and moreover the value will tend to vary with the difficulty of the item. It cannot be considered the equivalent of a Pearson correlation under any assumptions. This criticism of the Sampling principle is that the concept of correlation is not an appropriate means of expressing the relationship of an item to a criterion score.

The Sampling principle does not even pretend to offer a basis for selecting items when the goal is not predicting a criterion but improving the self-consistency of a test. In constructing tests for analysis by factorial methods, however, it is precisely self-consistency rather than prediction of a criterion which is the goal. Let us examine the Equivalence principle, which appears to be directed towards this goal.

Perhaps no one has ever come out with the flat statement, "You should select items which are equal in difficulty and correlate as highly as possible with each other." Certainly, however, it is agreed that we desire to have tests which are as reliable as possible. Kuder and Richardson (20) state, "It is implicit in all formulations of the reliability problem that reliability is the characteristic of a test possessed by virtue of the positive intercorrelations of the items composing it." Although they did not exactly recommend that all items of the test be of equal difficulty, they assumed that the difficulties were equal, and as already pointed out their formulas cannot result in perfect reliabilities except for tests in which all items are equally difficult. In an earlier article on item analysis Richardson (25) had made the same assumption of equal item difficulties. Dressel (5) appears to regard as a particular virtue of the Kuder-Richardson formula (20) the fact that it does measure the extent to which the items of the test conform to what has here been called the Equivalence principle, namely, all items being of equal difficulty and inter-item correlations being as near to unity as possible. Dressel's paper is probably as close as any one paper to being an advocacy of the Equivalence principle.

The two most widely quoted papers on the subject of the optimal difficulty of items are those of Symonds (29) and T. G. Thurstone (36). Symonds reached the conclusion, "The best test for measuring a typical school grade or class is a test in which all of the items have a difficulty such that they can be answered with fifty per cent accuracy by the average individual of the group." Symonds "proved", by a method more ingenious than rigorous, that the items should be diversified in difficulty only if the individuals to be discriminated were extremely heterogeneous in ability. "The best test designed to measure several consecutive grades or classes is one in which the items have been so selected that they range evenly in difficulty from the level of difficulty which can be done with fifty per cent accuracy by the average member of the lowest group to the level of difficulty which can be done with fifty per cent accuracy by the average member of the highest group to be tested."

Mrs. Thurstone also arrived at the conclusion that 50% was the level of optimal difficulty, but she did not recommend that all items be of this difficulty. Her

conclusions are conservative: "It seems fairly safe to guess that better differentiation between the abilities of a group can be obtained with a test in which the average percentage of error is about fifty per cent and in which the difficulty of the separate questions ranges from about thirty per cent to seventy per cent successes than can be obtained from a test in which the percentage of error is only from zero per cent to twenty per cent as is at present the case in most school examinations and mental tests."

The method of investigation used by Mrs. Thurstone was to correlate spelling "tests" made of items chosen in a restricted range of difficulty against the complete test consisting of 1000 items. In an extension of this type of research, Richardson (24) made up similarly a series of five "tests", each containing fifty items within a range of seventeen percentiles of difficulty (such as five to twenty-two per cent successes), and correlated each of the five tests against the criterion, 803 items from which the tests were chosen. Instead of correlating the tests against the criterion as it stood, Richardson considered the problem of selecting a certain per cent of the criterion scores as "passed" or "failed"; so he divided the criterion at various points and considered it as dichotomously scored. Biserial r was used as the measure of correlation. The correlations showed with striking clearness that each small test was most efficient at predicting the criterion when the proportion considered as passing the criterion corresponded closely with the proportions passing the items in the small test. The conclusions were equally clear: "If it is desired to separate off a minor proportion from the lower end of the distribution of criterion scores, then an easy test has much greater validity than have more difficult tests. Moreover, the smaller the minor proportion to be separated from the criterion group at the lower end, the easier should the test be. The converse situation applies to minor proportions to be marked off from the upper end of the criterion group." These results further "definitely point to the unsatisfactory nature of the common practice in the construction of tests of letting difficulty take care of itself." One further conclusion seems inescapable from these results, and it is very hard to see why Richardson did not draw this additional conclusion: In constructing tests of ability, where we are interested not in separating off a top group or a bottom group nor in dividing the group in the middle but in differentiating at all levels of ability, it is necessary to include items of all degrees of difficulty to get valid differentiation.

The two strongest arguments against the Equivalence principle of item selection are that a good test of ability is made up of items of all levels of difficulty, which Richardson proved in effect, and that the consummation of the Equivalence principle would result in a test giving results no different from those of a single item, which has been shown already as a consequence of Kuder and Richardson's derivations.

*B. Measurement of item validity.*

Whether items are selected according to the Sampling principle or the Equivalence principle or simply according to their correlation with a criterion, the problem arises of expressing the relationship between the dichotomously scored item and the continuously scored criterion or between one item and another. Probably the method of item selection most frequently employed is that of correlating the item with the original test of which it is a part, and selecting those

items with highest correlations. By an odd usage, the self-consistency of a test is referred to as reliability, while the consistency of an item with a test is referred to as item validity. While the correlation of a test with a criterion is almost always understood to mean the product-moment coefficient of correlation, there are dozens of ways of expressing the correlation of an item with a test. The choice of an index of item validity is thus a substantial one.

Most of the proposed indices of item validity are "thumbnail" statistics, and the majority of them are appreciably correlated with the difficulty of the item, as Long and Sandiford (22), in their comprehensive study, have shown. One would think that a most elementary requirement of a measurement of item validity is that it be independent of the difficulty of the item. Flanagan (7) has argued for this view, but Long (21) apologized for the fact that a coefficient which he developed was not correlated with difficulty, and immediately offered a modified form of the coefficient which would favor items of median difficulty. As the majority of these "thumbnail" indices are justified as shortcut ways of getting approximately the same information as biserial r affords, and as biserial r anyhow is used in many more studies than any other coefficient, it is of greatest historical importance.

Biserial r is the equivalent of Pearson r when the dichotomous variable represents two parts of normal distribution. Since the Pearson r has unity for its maximum value only when the two correlated distributions are of the same form, biserial r has a maximum value of unity only when the continuous variable is also normally distributed. The common practice of assuming that biserial r represents the definitive correlation between an item and its criterion is equivalent to saying that each item on a test represents a continuous range of some normally distributed ability, and that everyone who possesses more than a certain amount of the ability answers the item correctly, everyone who possesses less than that amount answers the item wrongly, and the ability thus defined has rectilinear regression on the criterion ability. There is an immediate intuitive validity to saying that the ability to do most test items is not an all or none phenomenon, that some people are just able to do a given problem, some people are able to do it easily, some people are not quite able to do it, and others are far from able to. There is not a corresponding intuitive justification for the statement that these differences in ability are normally distributed. Suppose that instead of having simply a right or wrong response on the given item, we had a whole set of scores on whatever ability the item measures. If this set of scores is not normal, then it can quite simply be changed to a normal distribution, by translating all scores into percentile scores and giving these percentiles their standard deviation values of the normal distribution. Since any set of scores on the given ability can be transformed to a normal distribution, it may seem not to be a large assumption to suppose that the ability represented by the right or wrong response to an item is normally distributed. The difficulty is that any set of scores can be transformed to non-normal distribution quite as easily, and the assumption of a normal distribution of ability is part of the very meaning of a biserial coefficient of correlation. Furthermore, there does not appear to be the remotest possibility of ascertaining the validity of the assumption that the hypothetical, normally distributed ability for

that item has a rectilinear relation to the criterion ability. But if this assumption falls down, there is no basis for considering biserial r as the definitive correlation between an item and its criterion. Besides the theoretical difficulty in assuming that the scores which are available only as zero or one actually represent a continuous, normally distributed variable with rectilinear regression on the criterion variable, there is a major practical objection to biserial r. The practical objection is the further assumption that the continuous variable is normally distributed. This assumption is of a different sort from the assumptions concerning the item variable. The assumptions concerning the item variable are completely inaccessible; if there is any way of determining their reasonableness, it has yet to be suggested. The assumption concerning the test distribution, which might better be called a restriction, can be tested from the data available before computing biserial r. When the assumption is not consistent with the data, biserial r definitely loses its justification as the equivalent of Pearson r, and in fact may assume values substantially greater than unity. Judging by the large number of reported biserial coefficients greater than unity, for example in Richardson's (24) paper, this restriction on the use of biserial r is by no means taken as seriously as it must be for this coefficient to have a claim to preeminence over the many "thumbnail" coefficients.

Calling attention to the inappropriateness of the assumptions underlying biserial r for the correlation of an item with a test, Richardson and Stalnaker (26) proposed another coefficient which has since become known as point biserial r. Point biserial r makes no assumption of continuity or normality, they claim, and has "the essential characteristics of the bi-serial r, and none of its disadvantages from the standpoint of underlying assumptions." Point biserial r is immediately seen to equal biserial r times $z/pq$, where $p$ is the proportion passing the item, $q = 1 - p$, and $z$ is the ordinate of the normal curve corresponding to the split $p$, $q$. The derivation of the coefficient was very simple; the formula for Pearson rectilinear regression was applied to the scatter-diagram of test against item. Their derivation is equivalent simply to applying the formula for product-moment correlation without regard for the restriction of this formula to many-valued variables. In discussing the rationale of item analysis, Richardson (25) began with the same formula, stating that other indices of item validity "are substitutes for or approximations to the ordinary coefficient of correlation between the item and the total test score." As has been stated, the use of point biserial r is implied by many discussions, such as the selection of items by the multiple regression technique.

Since point biserial r is a function of biserial r and the proportion passing the item, and since the properties of biserial r are well known, the corresponding properties of the new coefficient can be obtained by use of tables of areas and ordinates of the normal curve. It is easy to verify that point biserial r is a fraction of biserial r, less than one, how much less depending on the proportion passing the item. Thus the new coefficient does not have one for a maximum value—except when the continuous variable has only two values—and it is not independent of the difficulty of the item. Since it is computed by ignoring the restrictions on Pearson r, it cannot be considered the equivalent of a Pearson r. There appears to be no reason for con-

sidering this coefficient as essentially superior to the numerous other "thumbnail" coefficients.

While no entirely satisfactory rationale has yet been offered for any index of item validity, it does not follow that no satisfactory index has been proposed. In Chapter V, Section B, a rationale will be offered in support of a slight modification of a coefficient proposed by Long (21) virtually without rationalization.

### *C. The correlation of two items.*

The problem of expressing the correlation between two items can be discussed in about the same terms as the problem of measuring item validity. The tetrachoric coefficient of correlation is based on assumptions corresponding to the assumptions of biserial r, which are just as objectionable in application to two items as to an item and a test. The tetrachoric r does, however, have the virtues of equivalence to Pearson r when its assumptions are fulfilled, of a maximum value of unity, and if the appropriate formula is used, of independence of the difficulties of the two items. This coefficient is seldom used, however.

Most discussions referring to the correlations of items with each other, such as those of Kuder and Richardson (20) and Jackson and Ferguson (17), mean the coefficient which has been called the four-point r. Like point biserial r, this coefficient is computed by using the formula for Pearson r, ignoring the restriction of this formula to many-valued variables. The maximum value of the four-point correlation is unity only when the same proportion passes both items, and the maximum decreases sharply with increasing discrepancy in proportions passing the two items. Jackson and Ferguson themselves pointed out this fact in commenting on Kuder and Richardson's paper.

No doubt there are again many "thumbnail" statistics for expressing the relationship between two items. The two most widely accepted coefficients are unsatisfactory in one case for depending on assumptions which are entirely gratuitous in this context, in the other case for depending to a marked extent on the agreement in difficulty of the items, which ought to be measured separately. In Chapter V, Section C, an exceedingly simple coefficient will be proposed as a measure of the relationship of two items, independent of their relative difficulties and not dependent for its meaning on any assumptions about the item variables.

## Chapter III

## The Problem of the Unit of Measurement

*A. Methods of scaling intelligence tests.*

Thorndike (31) has stated that one of the three fundamental defects in contemporary measures of intelligence is that "how far it is proper to add, subtract, multiply, divide, and compute ratios with the measures obtained is not known." (p. 1) One of the main ways in which measurement of intelligence needs to be improved is devising a scale "on which zero will represent just not any of the ability in question, and 1, 2, 3, 4, and so on will represent amounts increasing by a constant difference." (p. 4) Flanagan (6) has stated, "Perhaps the most fundamental problem in developing a system of scores is the selection of the units by which the position along the trait continuum shall be expressed." (p. 3) While numerous methods of devising units for psychological tests have been proposed, and still more methods have been used on one or more tests, the discussion here will be focussed on three of the best known methods of scaling, the method Thorndike (31) used in constructing the CAVD test, Thurstone's (32) method of "absolute scaling", and the method, described by Flanagan (6), being used by the Cooperative Test Service.

The first of these methods to be explicitly described is that of Thurstone. Briefly, Thurstone's scaling method is as follows. When a test is given to a single group, the percentage passing any item is translated into an abscissal value of the normal curve and considered the scale value of the item. When two or more age or grade groups are tested, and the method is intended for this case, the scale values of items in one group are assumed to be perfectly correlated with the scale values of overlapping items in the other group. Overlapping items are those which are not passed by nearly all nor failed by nearly all of either group. If the scale values are perfectly correlated, then the scale values for one group can be translated into corresponding values in the other group by the ordinary regression equation, and the constants of the equation will be the parameters, means and standard deviations, of the two groups. In practice, of course, the correlation will be less than perfect, and an improved estimate of the scale value of an item is obtained by expressing all determinations of its scale value in terms of a single basic group and averaging these determinations.

Approximately perfect correlation between the scale values of overlapping items for two groups is a restriction on Thurstone's method, that is, it is both an assumption and a criterion for the use of the method. Using Burt's data for Binet test results on 3000 London school children, Thurstone (32) found the assumption satisfactorily fulfilled, although of course this instance is far from a demonstration of the universal validity of the method. Referring to the use by Trabue of a scaling method proposed by Thorndike in 1916, Thurstone (33) stated, "The method of absolute scaling may be considered as an improvement on Thorndike's scaling in two regards, namely, (1) by providing for the varying dispersion, and (2) by providing a rational procedure by which all of the data may be adequately taken into account. We have called the method absolute, not

in the sense of measurement from an absolute origin but in the sense that the scale is independent of the unit selected for the raw scores and of the shape of the distribution of raw scores."

Thomson (30) has pointed out that Thorndike's (31) publication of the CAVD test and of its method of construction was about simultaneous with Thurstone's publication of the method of absolute scaling, and that "Thorndike, without actually writing any equations, used exactly the same two principles as are explicit in [Thurstone's] equations in making his CAVD scales. He falls behind Thurstone, it is true, in two respects: he does not explicitly use the criterion of almost perfect correlation between the sigma-values in two groups (though he implies it), and . . . he uses the *range* of the sigma-values instead of their standard deviation" in certain equations. Further details of Thorndike's method are not essential to the present discussion.

The methods of Thorndike and Thurstone, are, then, similar in a number of respects. Both methods are based on an assumption that the ability measured by the test is normally distributed for various age or grade groups separately. Both methods allow for different variability at different levels or for different groups. Both are methods of scaling items rather than scores. Thurstone, at least in the two articles quoted above, does not give any method for deriving scores from the scaled values of the items. Thorndike (p. 408) recommends that tests be made up in levels with a definite number of items in each level, that the levels be about equally spaced in terms of scale value, and that the score be computed as number of items right.

The procedure described by Flanagan (6) for use in scaling the Cooperative Achievement tests appears to depend on about the same assumptions as Thurstone's method. The main difference is that Flanagan's is a method for scaling scores rather than items. The main assumption is that the distributions of groups differing in mean and standard deviation can all be made normal simultaneously. His method is to administer the test to several groups whose scores overlap but differ in mean and standard deviation. One of these groups is chosen as the basic group. Within each group, scores are normalized, that is, translated into percentiles and thence into abscissal values for the normal curve. The median of the basic group corresponds to some raw score which has a scale value in each of the other groups; the medians of each of the other groups correspond to raw scores which have scale values in the basic group. Any median has of course a scale value of zero in its own group. These scale values thus provide two estimates of the distance between each median and the median of the basic group, one estimate in terms of the standard deviation of the basic group, one in terms of the standard deviation of the other group. The ratio of these two estimates provides an estimate of the ratio of the standard deviations of the two groups. The method of then finding scale values for various raw scores is rather complicated, but there appear to be steps which correspond to Thurstone's assumption that the scale value of overlapping scores in two distributions will be perfectly correlated, and to Thurstone's averaging of the different determinations of the scale value of a single score, though the weights attached in the averaging process are not explicitly stated nor rationally determined in Flanagan's method. Flanagan then plots the distributions of the groups in terms of the new scaled scores and repeats the procedure, treating the

scaled scores as the raw scores were treated at first. The procedure is repeated until repetition does not effect any change in the means and standard deviations.

Flanagan presents an example to show that although his method and Thurstone's give somewhat similar results, the results are far from identical. In comparing the methods Flanagan says, "The present approach to a greater extent dodges the issue of theoretical justification for these units as 'absolute' units, but stresses the practical end of obtaining a set of stable and usable *basic units*. It should be noted that the final *basic units* of the present procedure do not necessarily provide precisely normal distributions in all of the groups used in the scaling. They produce a combined distribution which has frequencies below the selected points which are precisely those which would be obtained if each of the scaling groups were normally distributed with the means and standard deviations as specified." (p. 22)

*B. The rationale of methods of scaling.*

Despite the similarity of the scaling methods of Thorndike and Thurstone, they have very different conceptions of what they are doing. Thorndike's purpose in scaling CAVD intelligence, and in determining absolute zero in CAVD intelligence by a method not discussed here, apparently was to obtain "truly equal units", in the sense in which inches on a yardstick are truly equal. He concludes, "We have an approximate scale of intellectual difficulty from an absolute zero in equal units. . . . This scale is at all points more accurate than the best scales previously available; and is accurate enough for many scientific and practical uses from I to Q, covering the interval from the upper extreme of the feebleminded to the 98 or 99 percentile adult intellect." (p. 471) Thurstone, on the other hand, appears to claim no more for "absolute scaling" than that "the scale is independent of the unit selected for the raw scores and of the shape of the distribution of the raw scores." Exactly what is meant by "the unit selected for the raw scores" is not too clear, but Thurstone seems to be claiming that the scale is independent of the particular items included in the original test and of the particular groups used in scaling. Flanagan frankly avoids any elaborate rationalization of his procedure.

All three methods depend on the assumption that the ability measured by the test being scaled is normally distributed in the age or grade groups used. Flanagan again admits the assumption of normality is simply convenient, but a passage which might be considered justification for the assumption states, "An interesting characteristic, which increases the usefulness of units which have been derived to produce 'normal' distributions, is the relative independence of these units of the particular groups used in the derivation procedure." (p. 8) He does not offer any evidence in favor of this assertion, nor does he present internal evidence in his own data.

Thurstone does not claim that the true distribution of intelligence is normal, but only makes such claims as, "The application of the present method of scaling to Binet test data shows that the distributions of intelligence for children can be assumed to be normal at least as far up as the age of 14." (32) Actually, all that was shown directly by his criterion was that the scale values of the items, assigned on the assumption of normality in each group, were consistent in different groups. He did not show that the scores derived from the scaled values of

the items were normally distributed, nor did he give any hint as to how the scale values of the items would affect scoring. In later sections of this paper it will be shown that if a test is perfectly homogeneous, in a sense to be defined, scaling of items and scaling of scores are identical processes. The Binet-type tests are not, however, homogeneous in this sense.

Thorndike offers very elaborate proof that the distribution of intellect is truly normal. The general line of argument can be illustrated as follows. Curves for the distributions of various sixth grade groups on various tests of intelligence are shown to be of a variety of forms. "We are not concerned, however, with the form of distribution based upon any single test or examination. The form of any such single distribution, granting that the sample was both representative and numerically adequate, might not reflect the true form of distribution of intellect in this grade, either as the result of the error of measurement in the individual scores, or through the effect of inequalities in the units of the tests. In so far as inequalities in the units of the tests occur purely by chance, however, inequalities in one direction in one test will tend to be balanced by like inequalities in the opposite direction in some other test. We have therefore combined the eleven separate distributions, equal weight being attached to each, into a single composite distribution, by averaging the frequencies for each successive one-tenth sigma and plotting the resulting curve." (p. 522) The resultant average curve is of course a very close fit to the normal curve. Thorndike admits a flaw in the argument. "The normal curve bears an excellent reputation in psychological literature. One might conjecture with some show of reason, therefore, that in the construction of these tests there has been a more or less general and conscious effort to adjust the units of the tests so as to distribute the pupils according to the normal curve, and that since the sixth grade approximates the mean of the range of ability for which the tests have been generally devised, such deliberate inequalities would probably be most effective in and near the sixth grade." (p. 527) Using the same eleven tests as were used for the sixth grade, an average curve for eleven different ninth grade groups was shown also to be very close to the normal curve. Thorndike argues that it would be very difficult to introduce spurious inequalities in the unit of measurement so as to produce normal distributions in the ninth grade and in the sixth grade simultaneously. A similar argument, but different tests, are used to prove that intelligence is normally distributed among twelfth grade students and among college freshmen. Thorndike's point of view throughout is that there exists a true distribution of intelligence, which is waiting to be discovered, which may be distorted somewhat by a particular test, but in general can be discovered by means of existing tests. There is an odd contradiction, however, in Thorndike being so much of a specifist that he dare not name his test a test of intelligence, but had to name the kinds of intelligence in terms of the forms of the items, and yet being so little of a specifist that he assumed that the traits measured by a large variety of intelligence tests, using many other kinds of items and subject matter, are so nearly the same that the form of distribution of all of them must be identical for a single grade. The main argument does not come from the forms of the distributions for single tests, and these are by no means all good fits to the normal curve. The main argument is from the crude averag-

ing of distributions, and this argument is completely inadmissible. The tests simply are not all measuring the same thing, and thus the "true" distributions, at least in the sampling sense, need not be the same for all tests. If a number of distributions, no matter what traits they measured, were chosen at random from a series of statistics books, it would be found that some of the distributions were skewed to the right and some skewed to the left, some were platykurtic, and some leptokurtic. If a sufficiently large number of such randomly chosen distributions were averaged, any degree of goodness of fit to the normal distribution could be obtained, simply because the normal curve has "average" characteristics. Thorndike's main argument is purely a statistical artifact.

More important than the fact that Thorndike did not really prove anything about the distribution of intelligence when measured in truly equal units is the fact that the truly equal units which Thorndike and others have looked for are purely chimerical. A careful logical argument to this effect can be found in a recent paper by Bergmann and Spence (1). The main point of their argument can be summarized briefly. The properties of addition and subtraction, which are the characteristics by which Thorndike identifies "equal units", are essentially the properties not of scales but of the traits being measured. It is impossible to define an additive scale for measuring a trait for which the property of addition is not defined. Length is an example of a trait for which addition is defined. If we add one length to another length, we obtain a length, which is greater than either of the original ones. There simply is no way to add one intelligence to another intelligence to obtain a greater intelligence. Since intelligence cannot be added, there is no real meaning to ascribing "additivity" to a scale of intelligence, and certainly no test of whether any scale of intelligence has the property of "additivity". For traits for which addition is not defined, Bergmann and Spence continue, "We choose our scales so that certain empirical laws receive an expression as intuitive and (or) as mathematically simple as possible. It is an inaccurate and misleading way of speaking when such choice of scales is described as an attempt to equalize the unit distances of the scale." Since psychologists mostly use intelligence test scores for correlations, and in particular factor analysis is based on the correlations between tests, the stability of correlations provides a possible criterion for the choice of scales. This point will be elaborated in Chapter VI, where it will be made clear that this is essentially the criterion which Thurstone used.

There is an important covert assumption underlying probably any attempt at scaling psychological tests, certainly such methods as have been discussed here. Thomson (30) has stated clearly that the methods of Thorndike and of Thurstone depend on the assumption that the tests, or items, can be placed at points on a scale which have a fixed relation, regardless of the age group being tested. This can only be achieved if what is being measured is either all one ability or composed of various abilities in constantly weighted proportions. "For suppose that what is being tested depends on two or more factors, say $p$, $q$, etc., that the different tests depend in different proportions on $p$ and $q$, and that these mature at different rates. Then in one age group the order of difficulty of the tests may mainly be decided by their $p$-saturations, if we imagine the group to be immature in $q$, while in a later group the $q$-saturations

may be the preponderant influence; and thus two tests, if they differ in their proportionate composition in $p$ and $q$, may even be inverted in order of difficulty. The assumption, that is to say, is that what is being tested is either structureless, or that, if it have structure, this cannot be detected because it never changes from test to test or from age to age." What Thomson considers an assumption about the nature of intelligence will be interpreted in Chapter VI, Section A, merely as an assumption about the nature of the abilities being measured by a test suitable for scaling. Thorndike, again speaking in terms of "intellect" rather than in terms of whatever ability may be measured by a particular test, has formulated something very close to what will be considered the criterion for a test satisfactory for the purpose of scaling: "If each of the tasks, the number of which measures width, is perfectly intellectual, depending for success upon all of intellect and nothing but intellect, the change from one hundred per cent of successes to zero per cent of successes, as the intellect in question is tested at higher and higher altitudes, will be instantaneous. When a small amount of inadequacy and error is present, as in our 40-composites for Intellect CAVD, the change will still be very sudden." (p. 374) Thorndike, unfortunately, made no further use of this idea. His proof of the homogeneity of what is measured by the CAVD test was the high correlation between scores on successive levels. If it were true that individuals scored either 100 per cent or zero at any level, or close to those percentages, then correlation would be a poor technique for showing this fact. But the passage just quoted from Thorndike may very well have been germinal to the development of the statistical techniques which will be presented in the next chapter.

# PART II. OUTLINE FOR A SYSTEM OF CONSTRUCTING AND EVALUATING TESTS OF ABILITY

## CHAPTER IV

## THE HOMOGENEITY OF A TEST

*A. The concept of homogeneity.*

THE current solutions of the methodological problems in test construction and evaluation involve definitions and assumptions that are objectionable for various reasons. Is it possible to propose solutions for the same or similar problems on the basis of more satisfactory definitions and assumptions? This paper is intended as a contribution towards such solutions. The discussion will be limited to a specific class of tests: power tests of ability. There are many very important types of tests which are not included in this group, so that it may seem a disadvantage so narrowly to restrict the range of application. On the other hand, the restriction of the discussion has the advantage of enabling us to stay closer to the actual data, to construct definitions and assumptions which apply to the exact situation. Perhaps the basic difficulty in the currently used methods of test construction and evaluation is that these methods were imported from other fields where they were exact solutions, and were called on to function over a much wider range than that for which they possessed validity.

Let us begin by restricting the discussion to tests of ability, with every person having an opportunity to attempt every item, the score being number of items correctly completed. Ability will be taken to mean the immediate possibility of achievement. Admittedly there is no point in testing unless one intends to make inferences concerning the possibility of other achievements or the possibility of achievement under other circumstances. But the "tests of ability" with which the present discussion is concerned refer, in the first instance, to ability in the narrower meaning.

Two assumptions must be made: First, it is possible to test the same ability at different levels of difficulty. Second, for any two items in the same test, the possession of abilities required to do one item may help or may not help but they will not hinder or make less likely adequate performance on the other item. The second assumption need not hold for abilities in general but only for those closely enough related so that they will be involved in the items of one test. The first assumption does not explicitly enter the derivations which will follow. If the first assumption is wrong, the mistake does not invalidate the derivations, but it makes impossible the achievement of the goal with which the discussion is concerned.

In defense of the second assumption, the several decades in which abilities have been measured have yielded remarkably few instances of negative correlation of abilities. The first assumption certainly is implied in the notion of "factor score" and in the attempt to construct pure tests of the factors discovered in factor analysis. The believers in a general factor and the believers in group factors will find this assumption congenial; those who believe that the amount of intelligence can be reduced to the number of specific little things which one can do, adherents of some

form of the sampling theory of intelligence, probably will reject the first assumption. While the members of the sampling school of intelligence will have no use for the methods to be developed here, the methods will be equally useful for those striving to test a general factor and those testing group factors.

*Definition:* A *perfectly homogeneous* test of an ability is a test such that, if A's score is greater than B's score, then A has more of some ability than B, and it is the same ability for all individuals A and B who may be selected.

*Definition:* A *perfectly heterogeneous* test is a test composed of items each of which measures an ability independent of the abilities measured by the other items.

One might ask, why must the items measure independent abilities? Why is it not sufficient that the items measure different abilities? Let us imagine a perfectly homogeneous test consisting of a number of items. One more item is added. If the added item tests the same ability as the original test, the test will still be homogeneous; if it tests a closely related ability, the test will depart from homogeneity only slightly; but if it tests a completely independent ability, the test will depart from homogeneity still further.

*Theorem I:* When the items of a perfectly homogeneous test are arranged in order of increasing difficulty, every individual will pass all items up to a certain point and fail all subsequent items.

Proof: By the above definitions, every individual has more ability than individuals with lower scores, and it is the same ability, and ability is the immediate possibility of achievement. Therefore, every individual must be able to do correctly all problems done correctly by anyone with a lower score. In particular, he must be able to do correctly all problems done by someone with a score one point lower than his. A person with a score of, say, 37 can do correctly all problems done by a person with a score of 36, plus one more; so a score of 37 means that the 37th item, when the items are ordered according to difficulty, is the hardest item successfully completed, that all previous items are done correctly, and thus that all succeeding items are failed. This establishes the theorem.

A simple consequence is that two individuals with the same score must have completed successfully the same items, though they do not necessarily have the same amount of ability. Two individuals with different levels of ability may obtain the same score if there is no item in exactly the range of ability where one can succeed and the other cannot.

Comparing the concept of homogeneity to that of reliability, we find that it is closest to the split-half reliability, since homogeneity is decreased mainly by what were called accidental content factors in the discussion of reliability. Transitory variations in efficiency affect homogeneity only insofar as they are tied to specific, that is, accidental, content. Abilities which are specific to a few items of a test will show up as sources of heterogeneity as much as any other accidental content factors.

It could probably be shown that for any available test it is not possible to arrange the items so that everyone will do correctly all items up to a certain point and none of the subsequent items. It is no criticism of the concept of homogeneity that no perfectly homogeneous test exists, nor would one reject the concept of reliability because there is no perfectly reliable test. The importance of Theorem I is that it provides a criterion of homogeneity, a set of simple

operations by means of which we can recognize a perfectly homogeneous test if one exists, while the concept of reliability is not tied unambiguously to any set of experimental operations, as shown in Chapter I.

While Theorem I provides a necessary condition for a perfectly homogeneous test, there are two objections to considering it as providing a sufficient condition. A test could satisfy the criterion of homogeneity if each item depended on the same composite of abilities; there is nothing in the criterion which enables one to distinguish a test of just one ability from a test of a constantly weighted composite of abilities. In the discussion which follows this distinction will not be important; moreover, there are other methods, the methods of factor analysis, which are appropriate to making such distinctions.

Suppose we had a test composed of ten items testing different abilities, one item at a level of difficulty appropriate to each grade from one through ten. The test is given to a group of ten students, including an average student at each grade level from one to ten. Very probably such a test for such a sample would satisfy the criterion of homogeneity, regardless of the relationship between the abilities measured by the separate items. Satisfying the criterion for homogeneity is not in itself a sufficient condition for the test to be homogeneous. It will be obvious, however, that such wide gaps in difficulty will not occur in ordinary tests of abilities, such as are used in vocational counselling clinics, for example. Furthermore, if we extend the sample to include not only typical students but a random group of students, there will soon be included some who can pass fifth grade arithmetic but fail fourth grade spelling, and thus reveal the heterogeneity of the test. It appears that if wide gaps in difficulty of items are ruled out, and if the sample used to test the homogeneity of a test is truly random, then as the size of the sample increases, it becomes increasingly improbable that a test which is not homogeneous with respect to one or a group of abilities will conform to the criterion of homogeneity.

### *B. The measurement of homogeneity.*

The definitions of perfectly homogeneous and perfectly heterogeneous tests can be restated in terms of probability. In a perfectly homogeneous test, when the items are arranged in the order of increasing difficulty, if any item is known to be passed, the probability is unity of passing all previous items. In a perfectly heterogeneous test, the probability of an individual passing a given item A is the same whether or not he is known already to have passed another item B. The second basic assumption of this paper can also be stated in these terms. For any test of ability, the probability of passing any item A for those known to have passed any other item B is not less than the probability of passing item A for those whose response to item B is not known.

Let us denote by $p_i$ the probability of passing the $i$th item, by $p_{ij}$ the probability of passing both the $i$th item and $j$th item, and by $p_{i/j}$ (to be read "p i given j") the probability of passing the $i$th item among those known to have passed the $j$th item. As the quantity $p_{i/j}$ is defined only for values of $j$ for which $p_j$ differs from zero, it will be convenient to assume that there is no item in the test which everyone fails.

By the above definition,

(1) $p_{i/j} = p_{ij}/p_j$.

The second assumption corresponds to

the equation:

(2) $p_{i/j} \geqq p_i$.

The definition of a perfectly heterogeneous test in terms of probability corresponds to the equation:

(3) $p_{i/j} = p_i$.

*Definition:* Item $j$ will be said to be more difficult than $i$ if $p_j$ is less than $p_i$.

If the items of the test are arranged in order of increasing difficulty, the definition of a perfectly homogeneous test in terms of probability corresponds to the equation:

(4) $p_{i/j} = 1$, for all $j$ greater than $i$.

Apparently the degree of homogeneity of a test depends on the values of the quantity $p_{i/j}$ for all pairs of items, and an adequate index of homogeneity must be based on all of the values. For each pair of items, $j$ greater than $i$, the quantity $p_{i/j}$ has a value between a lower limit of $p_i$ and an upper limit of unity. It is natural therefore to build an index of homogeneity around the quantities $p_{i/j} - p_i$. The probability, $p_i$, of succeeding on a given item is a characteristic of the entire population, but the probabilities $p_{i/j}$ are characteristics of groups varying in size, depending on the magnitude of $p_j$. It is therefore proposed that the index of homogeneity weight the quantities $p_{i/j} - p_i$ according to $p_j$. Consider now the sum,

$$(5) \qquad S = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} p_j(p_{i/j} - p_i).$$

By equation (3), $S$ will equal zero for a perfectly heterogeneous test, since each term will equal zero. By equation (2), $S$ will always be positive or zero, as each term will be positive or zero. Considering the item difficulties as fixed and the relationships between items as varying, the maximum value of $S$ will be attained when $p_{i/j}$ assumes its maximum value of unity for each pair of items. By equation (4), this is exactly the definition of a perfectly homogeneous test.

$$S_{max} = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} p_j(1 - p_i).$$

A reasonable set of formal requirements for a coefficient of homogeneity is that it should have the value zero for a perfectly heterogeneous test, a value of one for a perfectly homogeneous test, and values between zero and one for intermediate degrees of homogeneity. Such a coefficient is provided by the ratio of $S$ to its maximum value, $S_{max}$. Let us call this coefficient $H_t$.

$$(6) \qquad H_t = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} p_j(p_{i/j} - p_i) \Big/ \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} p_j(1 - p_i).$$

Undoubtedly there are many other ways of combining the same quantities into an index of homogeneity having the same formal properties. The coefficient $H_t$ will have an advantage in ease of computation over many of the possible indices, as we will now show.

*Theorem II:* The coefficient of homogeneity, $H_t$, is a linear function of the variance of the test, with the constants of the function defined by the difficulties of the items.

The proof will be undertaken in two steps. First it will be shown that the variance of a perfectly homogeneous test or of a perfectly heterogeneous test depends only on the difficulties of the items. Secondly it will be established that

$$H_t = \frac{V_x - V_{het}}{V_{hom} - V_{het}}, \tag{7}$$

where $V_x$ is the variance of the test, $V_{het}$ is the variance of a perfectly heterogeneous test with the same distribution of item difficulties, and $V_{hom}$ is the variance of a perfectly homogeneous test with the same distribution of item difficulties.

The score of any individual on the test can be expressed as the sum of his scores in the separate items, with each item having a score of zero or one:

$$X = x_1 + x_2 + \ldots + x_m,$$

where $X$ is the score on the test, $x_i$ is the score on the $i$th item, and the test has $m$ items. Using the formula for the variance of a sum, we obtain

$$V_x = \sum_{i=1}^{m} V_i + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} r_{ij}\sqrt{V_i V_j},$$

where $V_i$ is the variance of the $i$th item, and $r_{ij}$ is the straightforward product-moment correlation between the $i$th and the $j$th item. But $V_i$ is known to be simply $p_i q_i$, and $r_{ij}$ is equal to

$$(p_{ij} - p_i p_j)/\sqrt{V_i V_j}.$$

Substituting these values into the right-hand side, we have

$$\begin{aligned} V_x &= \sum_{i=1}^{m} p_i q_i + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} (p_{ij} - p_i p_j) \\ &= \sum_{i=1}^{m} p_i q_i + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} (p_j p_{i/j} - p_i p_j). \end{aligned} \tag{8}$$

The last form is obtained by substitution from equation (1).

To obtain the variance of a perfectly homogeneous test, we substitute from equation (4) into equation (8):

$$V_{hom} = \sum_{i=1}^{m} p_i q_i + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} (p_i - p_i p_j). \tag{9}$$

Similarly, substituting from equation (3), we obtain as the variance of a perfectly heterogeneous test,

$$\begin{aligned} V_{het} &= \sum_{i=1}^{m} p_i q_i + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} (p_j p_i - p_i p_j) \\ &= \sum_{i=1}^{m} p_i q_i. \end{aligned} \tag{10}$$

Substituting from (8), (9), and (10) into the right-hand side of (7), we obtain,

$$\frac{V_x - V_{het}}{V_{hom} - V_{het}} = \frac{\sum_{i=1}^{m} p_i q_i + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} (p_j p_{i/j} - p_i p_j) - \sum_{i=1}^{m} p_i q_i}{\sum_{i=1}^{m} p_i q_i + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} (p_i - p_i p_j) - \sum_{i=1}^{m} p_i q_i} = \frac{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} p_j (p_{i/j} - p_i)}{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} p_j (1 - p_i)}$$

Comparing the last form with equation (6), we see that this establishes equation (7). This completes the proof of Theorem II.

The coefficient of the homogeneity of a test, $H_t$, is defined in equation (6) in terms of the probabilities of passing successive items and the probabilities of passing the easier of two items granted that the harder of the two is passed, for all pairs of items. Theorem II shows that $H_t$ can be expressed in terms of the population variance and the probabilities of passing successive items. These probabilities are simply proportions of the population. The problem now arises of estimating $H_t$ in terms of character-

istics of a sample. The simplest way of forming such an estimate is to substitute the sample variance, $V_{sample}$, for the population variance, $V_g$, and the proportions passing successive items in the sample, $P_i$, for the probabilities, $p_i$. As the true order of difficulty can only be ascertained from the population, the order, $i$, must also be estimated from the sample. Making these substitutions, and a number of algebraic simplifications, the following computational form for the estimate of homogeneity, *Est* $H_t$, can be derived:

$$\text{Est } H_t = \frac{N(\sum X_k^2 - \sum X_k) + \sum N_i^2 - (\sum X_k)^2}{2N(\sum iN_i - \sum X_k) + \sum N_i^2 - (\sum X_k)^2}, \tag{11}$$

where $X$ refers to raw scores, $N_i$ refers to the number passing the $i$th item, when the items are ordered according to decreasing number passing, the subscript $k$ means summation for all $N$ individuals, and the subscript $i$ means summation for all $m$ items. The derivation of formula (11) is presented in an Appendix.

The argument for the use of the coefficient of homogeneity derived here is that it describes an important characteristic of tests of ability. But there are many situations where it is imperative to adopt descriptive statistics which are economical in terms of computation. Inspection of formula (11) shows that, aside from the tabulation of the number passing each item, the computation of *Est* $H_t$ requires very little more work than the computation of two standard deviations.

The problem of the sampling properties of the estimate of homogeneity is an important one, but these properties have not been ascertained as yet. It will be of interest to know the magnitude of the expected sampling error under various conditions, and of even greater interest to know whether the proposed estimate is unbiased. It may very well be the case that some other quantity will provide a better estimate, particularly for small samples. Until such time as these questions have been investigated, it seems advisable to use the proposed estimate of homogeneity only for fairly large samples, say, over 100 cases.

## Chapter V

## The Homogeneity of an Item with a Test

### *A. The principle of item selection.*

The foregoing discussion of the homogeneity of a test contains implicitly the aim of constructing tests as homogeneous as possible. A test departs from homogeneity when individuals pass items above the levels of difficulty defined by their total scores and fail items below their levels of difficulty. The items may very well differ in the extent to which the individuals who pass are the ones who would be expected to on the basis of total score. There seems to be no reason why the consistency of the item with the test should not be described with the same term as the internal consistency of the test, namely, homogeneity.

*Definition:* An item will be said to be *perfectly homogeneous* with a test if all of those passing the item have higher scores on the test than all of those failing the item.

*Definition:* An item will be said to be *perfectly heterogeneous* with a test if the scores on the test of those who pass the item are distributed at random with respect to the scores of those who fail the item, when the scores are ordered according to magnitude.

It is obvious that the test will be perfectly homogeneous only if each item is perfectly homogeneous with the test. In a perfectly homogeneous test two individuals will be discriminated from each other only if there is at least one item at the level of difficulty where one can succeed and the other cannot.

The following principle is therefore proposed as a rational principle of item selection, consistent with the aim of forming perfectly homogeneous tests: Each item should be as homogeneous with the total test as possible, and there should be one or more items at every level of difficulty which the test is intended to cover.

In practice, the safest rule would be to include items which nearly every one passes and items which nearly every one fails. Exactly how the item difficulties should be distributed between these extremes must be determined by other considerations. There is no way of telling whether an item which everyone passes or everyone fails is testing the same abilities as the other items; so in order to be sure that the items extreme in difficulty are testing the same abilities as the rest of the test, the group used in standardization must be more variable than the groups among whom the test will be expected to discriminate later.

In a perfectly homogeneous test one item at any level of difficulty is all that is required, since another item at the identical level of difficulty will give the same information as the first one. One might draw from this fact a second principle: There should not be two items at any level of difficulty in one test. As no perfectly homogeneous test ever will be constructed, this rule is less important than the first, but where the aim is to get the maximum information with the minimum number of questions, it is a useful principle: We may expect that the effect of the residual heterogeneity will be minimized by continuing to assign scores according to the number of correct items, rather than according to the ordinal number of the hardest item done correctly, a theoretical alternative.

### *B. Measurement of the homogeneity of an item with a test.*

Perhaps many of the indices of item

validity now extant would give a rating close to one to the items which have been defined as perfectly homogeneous with the test. It is not so clear that an item which is perfectly heterogeneous with the test would receive a rating of zero by all of these methods, but perhaps the rating would be very low. In selecting or constructing an index of the homogeneity of an item with a test, two requirements must be met: Since, by the principle of item selection adopted, it is desirable to have items in all ranges of difficulty, the index of homogeneity should not be prejudiced by the difficulty of the item. And since we cannot assume the problem of scaling to have been solved, the index of homogeneity should not be prejudiced by the unit of measurement of the test. In other words, the index must utilize only the order, not the magnitude, of total scores.

An index proposed by Long (21) has essentially these properties. The logic of this index can be explained as follows. If the number of people passing the item be indicated by $P$, and the number failing be denoted by $Q$, then the item makes a total of $P$ times $Q$ discriminations. For each person who passes the item is judged by that item to be better than each person who fails the item. Using the total score as criterion, if each of these discriminations is made correctly, all those who pass will have higher scores than all those who fail. For every pair of persons such that the one who passes has a lower score than the one who fails, the item has made one wrong discrimination. If we divide the total number of wrong discriminations by the total number of discriminations made, the result is the percentage of wrong discriminations. Long's index, which he calls an index of overlapping, is constructed by subtracting twice the percentage of wrong discriminations from one. In computing, one counts for each person failing the item the number of those with lower scores who pass the item and sums the quantities for all those failing a given item.

$$\text{L.O.} = 1 - \frac{2\sum \text{"passes" below "fails"}}{PQ}.$$

Since the proportion of wrong discriminations can assume values from zero to one, the index of overlapping will assume values from minus one to plus one. The value of plus one will be assumed when all discriminations are correct, and the value of minus one when all discriminations are wrong.

It is clear that Long's index utilizes only the order, not the magnitude of total scores. It appears to be unrelated to difficulty, and in an experimental study (22) it has been found to be approximately uncorrelated with difficulty. There are, however, a few objections to adopting this as an index of item homogeneity as it stands. Long did not mention the possibility that two people might be tied for total score, with one passing and one failing the item. But obviously as soon as the number of people is greater than the number of items, there will be a considerable number of ties, and if the test is not perfectly homogeneous, we may well expect that some of those with tied scores will differ in respect to particular items. The decision as to handling ties therefore will have a sizeable effect on the magnitude of the index in some cases. The second problem is whether the item should be counted in the total score. The third problem, and the hardest one, is whether the index will have an expected value of zero in the case of a chance relation of the item and the total score.

Consider first two persons whose total scores, including the item, are identical but only one of whom passes the item. On a test which is the same as the original test but omits the item being evaluated, the individual who has failed the item will get the same total score he now has, but the individual who passes the item will have a total one point lower than his original score. Thus, to eliminate a spurious positive correlation between item and total score, we must count all pairs of persons who are tied for total score but who differ in item score as "wrong discriminations" on the part of the item. But now consider two people such that one has a score one point higher than the other but the person with the higher score passes the item while the person with the lower score fails the item. If we omit the item from the total, the higher score will be reduced by one point, while the lower score will stay the same, so these two people will now be tied. In deciding how to count this case, it should be remembered that in a perfectly homogeneous test it will happen regularly that there are two individuals who are discriminated by just one item of the test, the item at exactly the degree of difficulty where one can just pass and the other cannot quite pass. If there is only one item at this level of difficulty, the omission of this item will make these individuals tied for total score, but this does not mean that the item discriminated between them wrongly. Suppose we agree, therefore, that when two people are tied on the total minus the item, we shall not in any case count them as wrongly discriminated. Notice that for two such people, it does not matter which one passes the item and which one fails, since in either case it is a "correct" discrimination, according to the agreed way of counting. If the number of tied scores is very great, the item is automatically credited with a considerable number of "correct" discriminations by this method, and the "percentage of wrong discriminations" will assume values from zero to a fraction less than one. In order to make the percentage of wrong discriminations a true percentage, we shall have to eliminate the tied pairs both from the numerator and the denominator. On the basis of the above considerations it is proposed that Long's coefficient be modified so that "passes" tied with "fails" (on the total including the item) be counted in the numerator and "passes" one point above "fails" be subtracted from the denominator of the percentage of wrong discriminations. Let us call this new coefficient the homogeneity of the item with the test and denote it by $H_{it}$.

$$H_{it}=1-\frac{2\sum \text{"passes" below or tied with "fails"}}{PQ-\sum \text{"passes" one above "fails"}}. \tag{12}$$

The coefficient $H_{it}$ takes care of two of the objections to Long's coefficient, but its adoption for actual use should be contingent on the demonstration that for items perfectly heterogeneous with the test, the expected value of the coefficient will be zero. Long perhaps assumed that since his coefficient varied from minus unity for perfect inverse relation to positive unity for perfect positive relation, it would be zero for no relation, but this conclusion does not follow. Since in the field of testing abilities we are interested mainly in degrees of positive relation, the requirement that the coefficient be minus unity for perfect

negative relation is not at all important; the value of the coefficient for chance relationship is a more meaningful reference point. An investigation of the sampling properties of the above coefficient is necessary to establish the value to be expected for a chance relation.

*C. Measurement of the homogeneity of two items.*

The relationship between two items in a perfectly homogeneous test reduces simply to this: All those who pass the harder item also pass the easier item. In the case of two items of identical difficulty, the same individuals pass both items. In terms of a four-fold table showing the relationship of the items, the condition means that at least one of the cells in the negative diagonal will have no entries. For items differing in difficulty, this cell will be the one showing those who pass the difficult item and fail the easy one; for items identical in difficulty, both cells of the negative diagonal will have zero entries. We will deal only with the case of items differing in difficulty, since for items identical in difficulty, the resultant coefficient will turn out to be the same whichever item is arbitrarily considered the more difficult.

Given fixed proportions passing and failing the two items, a single cell entry determines completely the fourfold table. A coefficient describing the relation of the two items can therefore be based entirely on the entry of one cell and the totals passing and failing the two items. It is proposed to use the cell which, under the condition of perfect homogeneity, would have an entry of zero. Every individual who passes the harder item and fails the easier indicates a discrepancy in classification according to the two items. The maximum number of such discrepancies is either the number who pass the difficult item or the number who fail the easy item, whichever number is smaller. It is easily seen that in this case the number of discrepancies to be expected if there is a chance relation between the items is not exactly half the maximum number of discrepancies, for, according to a chance relation, the expected number of discrepancies is the product of the proportions passing the harder times the proportion failing the easier times the total number of cases. It is proposed therefore that to measure the homogeneity or consistency of two items, we use the chance expectancy rather than the maximum as a base for expressing the percentage of discrepancies.

$$(13)\quad H_{ii}=1-\frac{K}{P_hQ_e/N}=1-\frac{NK}{P_hQ_e}.$$

where $H_{ii}$ is the coefficient of homogeneity of two items, $N$ is the number of cases, $P_h$ is the number passing the harder item, $Q_e$ is the number failing the easier item, and $K$ is the number passing the harder and failing the easier item.

By inspection, the coefficient $H_{ii}$ varies from unity for two items which are perfectly homogeneous to zero for two items which are unrelated. In general, it will not have the value of minus one for two items which are perfectly inversely related. This characteristic does not seem to be a disadvantage in the present problem, for while we do not exclude the possibility of two items having a negative relation, a negative relation of any magnitude is sufficient to prevent the two items from being included in the same test, and the main concern is with discriminating degrees of positive relation.

The coefficient of homogeneity of two items has been derived by logic some-

what similar to the logic of the coefficient of homogeneity of item with test. Both coefficients resemble a coefficient of rank correlation recently proposed independently by Kendall (19) and Rosander (27) to cover a somewhat different type of problem. They dealt with the problem of two paired sets of rank orders, excluding the possibility of ties in either test. The coefficient of correlation proposed involves using one set of ranks as criterion and counting the number of wrongly discriminated pairs of individuals in the other set. A wrongly discriminated pair of individuals is a pair such that one has a higher rank on one set and the other has a higher rank on the other set. Obviously it does not matter which set of ranks is used as criterion. There are $N(N-1)/2$ pairs of individuals in a group of $N$, since each one is paired with every other one. The coefficient proposed was

$$r = 1 - 2\frac{K}{N(N-1)/2} = 1 - \frac{4K}{N(N-1)},$$

where $K$ is the number of wrongly discriminated pairs. As the proportion of wrongly discriminated pairs can go from zero to unity, the coefficient can go from minus unity to plus unity. A consideration of the sampling distribution of the statistic showed that for chance relation between the two sets of ranks, the coefficient would have an expected value of zero. Unfortunately the coefficient as it stands is not usable even for correlating two tests, since unless the number of items is much greater than the number of people, ties in ranking will occur frequently. It seems fully possible that a more general coefficient of rank correlation could be worked out, permitting any number of ties, and including the correlation of item with test and item with item as special cases. The coefficients proposed above for these situations might or might not prove to be the equivalent values of this general coefficient of rank correlation. The difficulty is mainly not one of defining the coefficient but one of finding the corresponding sampling distribution.

## Chapter VI

## Criteria for an Adequate System of Scaling

*A. The problem of scaling.*

Psychologists generally admit that when someone sits down and constructs a test of mental ability, to be scored by counting the number of correct items, the resultant "unit of measurement" is an arbitrary one. The various attempts to replace these arbitrary scores by "equal units" or a rational set of scores are called the scaling of the test or scores. Before accepting any of the proposed methods of scaling, or proposing a new one, it seems desirable to state clearly what is meant by an arbitrary unit of measurement and to formulate rigorously the requirements for an adequate system of scaling.

There are various manifestations of what is called an arbitrary unit of measurement. One of the most essential and one susceptible of easy proof is stated in the following proposition.

*Theorem III.* The product-moment coefficient of correlation between two homogeneous tests of the same ability will in general be less than unity.

Proof: An example of two homogeneous tests of the same ability can be obtained by deleting an item from a homogeneous test, and calling the remainder of the items the second test. Suppose that the item is one which some but not all individuals could pass. All individuals whose scores were below the ordinal number of the deleted item when items are ordered according to difficulty, will have the same score on the second test as on the first, by the assumption of homogeneity. All individuals whose scores were at or above the ordinal number of the item will have their scores decreased by one. Thus the scores on the second test cannot be a linear function of scores in the first test. Since the condition of linear relation between two sets of scores is the necessary and sufficient condition for perfect product-moment correlation, the two tests will not have a correlation of unity.

It will be seen readily that there are any number of ways of deleting items from one test to form another test, and the correlation between the two tests can be changed radically by the choice of items for deletion. A rank order coefficient of correlation could be defined so as to permit ties in rank on one test without corresponding ties on the other test, by analogy with the coefficients proposed above for the homogeneity of an item with a test and the homogeneity of two items. Such a correlation coefficient would have the characteristic that it would equal unity for any two homogeneous tests of the same ability. The product-moment coefficient of correlation clearly does not have this characteristic. The problem of scaling may be thought of as the problem of defining an alternative set of scores to the original scores, such that two homogeneous tests of the same ability will have a product-moment correlation of unity.

We can now see that, in effect, the use of the coefficient of correlation in estimating the reliability of a test by the method of comparable forms or by the split-half method virtually assumes that the problem of scaling already has been solved. For the factors which decrease homogeneity are about the same as the factors which decrease split-half reliability, and these factors also decrease the comparable forms reliability. A perfectly homogeneous test is very much like a per-

fectly reliable test, at least by the split-half criterion of reliability. Theorem III could be reformulated, with less exactitude, to read that the correlation between two perfectly reliable tests of the same ability is in general less than unity.

The traditional solutions of the problem of reliability virtually assume that the problem of scaling has been solved; but the traditional solutions of the problem of scaling virtually assume that the problem of reliability has been solved. For the scaling of a test can be thought of as a method of refining the expression of differences between individuals, and if these differences are mainly matters of chance or non-systematic factors in the first place, no very refined information can be expected to result from the scaling process. Moreover, such methods of scaling as Thurstone's, defining the scale value of the items rather than the scores, give a set of scaled scores without the addition of further methods and assumptions only if there is a perfect correspondence between items and scores. The condition of perfect correspondence between items and scores is exactly the criterion chosen for a perfectly homogeneous test. Thurstone does not state whether he assumes such a correspondence, but neither does he mention the fact that there is a discrepancy between his scaled items and a set of scaled scores. If we think of these methods of scaling as assuming perfectly homogeneous tests, they are virtually assuming perfectly reliable tests, since there is a close relation between the concepts of homogeneity and reliability.

We may think of the process of measurement as taking place in two steps, first, defining the rank order of objects possessing the trait being measured and second, defining the differences between the amount of the trait at the various ranks. The rationale of aiming to construct perfectly homogeneous tests is that only with perfectly homogeneous tests does the number of items right yield a non-arbitrary set of ranks. Consider a test which does not satisfy the criterion for homogeneity. For example, person A passes only items 1, 2, and 3, while person B passes only items 3 and 4. As the test stands, A has a higher score than B, but if either item 1 or 2 is deleted, they will be tied. If, however, both item 1 and item 2 are deleted, B will rank higher than A. The same conclusion would hold if there were in addition 30 items, say, that both could do correctly. As every test-constructor knows, there is a considerable arbitrary element in the exact items chosen for inclusion in a particular test. The above illustration is sufficient to show that with a non-homogeneous test the arbitrary choice of items determines the rank order of the individuals, since the ranks can be reversed by simple deletion of items. With a perfectly homogeneous test, however, the deletion of items will result in an increased number of tied ranks, but it cannot reverse the rank order. In this sense, the criterion of homogeneity is the criterion for a non-arbitrary set of ranks.

The proposed measure of the homogeneity of a test and the considerations in selecting items to form a homogeneous test are based only on considerations of rank order. In the case of the measure of the homogeneity of the test, the use of the variance of the scores appears to admit the arbitrary differences between ranks, that is to say, the actual values of the scores. It must be remembered that the variance is considered only in relation to the distribution of item difficulties, which in effect cancels the arbitrary influence of the particular items included.

The traditional solutions to the prob-

lems of reliability and scaling are related in an unfortunately circular manner. The methods so far proposed for constructing homogeneous tests have not assumed a solution to the problem of scaling. But in discussing the problem of scaling in a precise and rigorous manner, it will be necessary to assume that the problem of constructing homogeneous tests has been solved. Scaling is essentially the second step in measurement, the step of assigning non-arbitrary magnitudes to the differences in rank. If the ranks are themselves arbitrary, it is hopeless to try to invent non-arbitrary magnitudes to represent the differences between successive ranks. This fact is the rationale of assuming perfectly homogeneous tests in stating the problem of scaling and in proposing criteria for its solution.

### *B. Criteria for an adequate system of scaling.*

Consider a perfectly homogeneous test T composed of $m$ items, and suppose the test is given to a group of individuals, call them sample A. Assume the test to be scored according to number of right items, and denote the possible scores by $x$; $x$ will clearly have the various integral values from zero to $m$. Suppose the items are ordered according to increasing difficulty, and denote the proportion passing the $i$th item by $P_{xAi}$, meaning, the proportion of sample A answering the item $i$ on the test whose scores are denoted by $x$. For short, let us denote the $m$ values of $P_{xAi}$ by $P_{xA's}$.

*Definition:* A *system of scaling* is a set of operations which defines a variable, $y(x_i, P_{xA's})$, a function of $x_i$ and of the $m$ quantities $P_{xAi}$, such that for all values of $x$,

$$y(x_i + 1, P_{xA's}) \geqq y(x_i, P_{xA's}),$$

and for some range of values of $x$,

$$y(x_i + 1, P_{xA's}) > y(x_i, P_{xA's}).$$

The function $y(x_i, P_{xA's})$ so defined will be called the derived scale for the test T, and the derived scale will be said to be *uniquely defined* in the range where $y$ increases with $x$.

In formulating criteria for an adequate system of scaling, claims made by the proponents of various systems will furnish valuable clues. Thorndike's claim that his method of scaling provides "equal units" or an additive scale is inadmissible simply because there is no criterion for an additive scale of mental ability. This argument was elaborated in Chapter III, Section B. Thurstone (33) has claimed that a scale derived by his method of "absolute scaling" is "independent of the unit selected for the raw scores and of the shape of the distribution of raw scores." Other writers have made similar claims for Thurstone's and for other systems of scaling. The meaning of the "unit selected for the raw scores" is not, however, entirely unambiguous. The "unit of measurement" in the case of a perfectly homogeneous test can be thought of as referring to the successive increments in difficulty of the items. The choice of the unit of measurement means the choice of the difficulties of the items, in other words, the choice of the specific items. In the case of a test which is not perfectly homogeneous, two people can get the same score by doing correctly entirely different items, and a person with a higher score will generally have failed certain items done correctly by persons with lower scores. Just what meaning can be attached to the term "unit of measurement" in this case is especially hard to say, but if it means anything, it again means the particular items chosen for inclusion in the test.

Thurstone's first claim seems to mean, then, that the "absolute scale" is independent of the particular test of the ability (or abilities) which is utilized in the scaling process. The second claim, that the derived scale is independent "of the shape of the distribution of raw scores", leads us to examine the meaning of independence. The basis for this claim appears to be the satisfaction of the criterion that the scaled values of the items in one age or grade group will correlate very highly with the scaled values in another age or grade group. Moreover, all the methods of scaling result in scales with two constants, one added and one multiplied, which are determined arbitrarily or by considerations other than those involved in the scaling process. These two constants will not affect the correlation of such a scale with any other variable, as they will not affect relative positions within the group. Independence thus seems to refer to relative position, that is, it means independence in the sense of correlation. In common sense terms, Thurstone appears to be claiming that the same scale of ability will result from his method of scaling, whether one starts with one test or another test of the same ability, and whether one starts with one age group or another. These claims correspond closely to the intuitive requirements for a non-arbitrary or rational scale of ability. Let us now formulate more exactly how these claims can be tested.

*Criterion I:* Given two homogeneous tests of the same ability T and T′, representing the scores corresponding to a given level of ability $i$ by $x_i$ and $u_i$ respectively, and representing the proportions passing successive items in sample A by $P_{xA's}$ for test T and by $P_{uA's}$ for test T′, then for an adequate system of scaling, $y(x_i, P_{xA's})$ will have a very high correlation with $y(u_i, P_{uA's})$ in the range where both derived scales are uniquely defined.

*Criterion II:* Given two samples, A and B, drawn from two populations within some broad class of populations, and representing the proportions passing the successive items in test T by $P_{xA's}$ for the first sample and by $P_{xB's}$ for the second sample, then for an adequate system of scaling, $y(x_i, P_{xA's})$ will have a very high correlation with $y(x_i, P_{xB's})$ in the range where both derived scales are uniquely defined.

These two criteria constitute a definition of an adequate system of scaling. In order to test whether the first criterion is satisfied, we need to have two tests known to be homogeneous tests of the same ability. These two tests are then given to a single group of individuals and the process of scaling is carried through. Each individual then has a scaled score on each test, and the correlation is computed between the two sets of scores. To test whether the second criterion is satisfied, the same test is given to two groups of individuals differing in average level of performance. Because the test is assumed to be homogeneous, to every score there corresponds an item, whose ordinal number is numerically the same as the score, and which is the hardest item which an individual with that score could do in that test. The process of assigning scale values to scores is therefore identical with assigning scale values to items. In this case, the correlation to be computed is that between the scale values of the items in one group and the scale values of the same items in the other group. Only those items are included which correspond to scores for which both derived scales are uniquely defined. No doubt these will be the items which are neither passed by nearly all

nor failed by nearly all of either group.

There is a major practical difficulty in testing for Criterion I, that is, how do we know whether two homogeneous tests are testing the same ability? A variety of possibilities suggest themselves. We may obtain one homogeneous test from another by deleting items. Or we may divide a homogeneous test into two tests, so long as there is considerable overlap in the difficulties of the items. An important consequence of satisfying Criterion I yields a secondary criterion which is easier to apply.

*Criterion IA:* A necessary but not sufficient condition for Criterion I to be satisfied is that the derived scale, $y(x_i, P_{xA's})$ shall not necessarily have a perfect correlation with the original scale, $x_i$.

*Proof:* As shown in Theorem III, two homogeneous tests of the same ability will not necessarily have a correlation of one, in terms of the original scores. If a system of scaling is proposed such that for both tests the scaled scores will correlate perfectly with the original scores, the scaled scores will correlate perfectly with each other only in the exceptional case where the raw scores show perfect correlation. Thus Criterion I will not be satisfied.

### *C. Are proposed methods of scaling "adequate"?*

We now have criteria in terms of which to evaluate proposals for scaling tests of ability. The inadequacy of percentile ranks and of standard scores as methods of scaling is immediately apparent. Percentile ranks depend entirely on the shape of the distribution for the sample; they pass Criterion IA, but they would generally fail Criterion II. Standard scores are linear functions of the original scores; they pass Criterion II but fail Criterion IA.

The criteria for an adequate system of scaling have been shown to correspond to the claims Thurstone makes for his method of "absolute scaling." Has Thurstone presented evidence sufficient to justify his claims? He has presented no direct evidence concerning Criterion I, but evidently the method of normalizing distributions does result in a scale which satisfies Criterion IA. Concerning Criterion II, he has not shown analytically that his method will generally result in the satisfaction of this criterion, and probably it is not possible to do so. He has presented evidence that in certain cases Criterion II is satisfied. If we rely on empirical proof, the evidence should be relevant and typical. The important problem is scaling scores and not scaling items, and his evidence concerns items. Scaling scores is identical with scaling items only in the case of homogeneous tests; moreover, there is little point in scaling unless one is using homogeneous tests. As Thurstone's evidence is in terms of tests which are almost certainly grossly heterogeneous, the evidence is not relevant enough.

A paper by Grossnickle (12), describing a method of scaling adapted from psychophysics, leads to a further precaution against the uncritical acceptance of empirical evidence in favor of a system of scaling. Miss Grossnickle presents evidence, though certainly not very elaborate evidence, that the scaled scores of individuals were more or less independent of the group in which they were included for purposes of scaling. The evidence was that the distance between individuals in one scale was an approximately linear function of the corresponding distance in the other scale. Both derived scales

turned out to be approximately linear functions of the raw scores, however. The high correlation between derived scores and raw scores may very well be accidental, as Miss Grossnickle believes, and therefore not a reason for rejecting her method of scaling. But the high correlation between the two sets of scaled scores using different groups may be a result of the fact that both scales correlate highly with the raw scores, and thus may also be an accident. In terms of Thurstone's method, if the raw score distributions are approximately normal, and they are then normalized, the resultant scale accidentally will be a linear function of the original scale. Two scales which accidentally happen to be linear functions of the raw scores by the same accident will be linear functions of each other.

It may be concluded that for empirical evidence of the high correlation of scales derived from two samples to be acceptable, it must be shown that one or both of the scales are not linear functions of the raw scores in that particular instance. On this basis, again, Thurstone's evidence in favor of his system of scaling is unsatisfactory.

The subject of scaling is far from closed. No system of scaling has been proved adequate by the criteria proposed here, though these criteria correspond to the claims made for Thurstone's system. Thurstone's method may yet be proved to be adequate, at least for some types of data. In any case, the development of adequately scaled tests awaits the development of highly homogeneous tests.

## Summary

Suppose that in constructing a test Psychologist Smith evaluates its reliability in terms of the split-half correlation, corrected by the Spearman-Brown formula. In improving the reliability of the test, he uses the biserial correlation of the item with the test to decide which items to keep and which to reject. After the final form of the test is determined, he scales the scores by normalizing them. This hypothetical psychologist may very well be a modal test constructor; at any rate he is not markedly atypical. Consider now the assumptions he has used at various steps.

In using an estimate of reliability based on the Spearman-Brown formula, he has committed himself to the assumptions that the error factor in the odd items is uncorrelated with the error factor in the even items, that the error factors in both halves are uncorrelated with the true scores in either half, that the error variance is identical in both halves, and that the "average" or "expected" performance of each subject differs from his "optimal" performance by the same amount as the average of any other subject differs from his optimal performance. There are additional assumptions about the equivalence of the two halves of the test; the above assumptions have been stressed because they also underlie such usages as the standard error of measurement, the usual proof that the reliability coefficient is equal to the ratio of the true variance to the obtained variance, and so on.

If we ask the hypothetical psychologist, "Exactly what is the characteristic of a test which you refer to as its reliability?" he will answer somewhat as follows: "Reliability is the correlation between two perfectly comparable forms of a test, and comparable forms are forms for which the true scores are equal, and true scores are the average scores on an infinite number of comparable forms administered with no effect of one testing on the next." Or he may answer: "Reliability is the ratio of the variance of true scores to the variance of obtained scores, and true scores are the average scores of an infinite number of parallel forms administered with no effect of one testing on the next, and parallel forms are tests which correlate to the extent of their reliabilities." At best he will answer: "Reliability is the correlation between two equally excellent measures of the same thing," but he will be unable to tell us further what he means.

In using biserial r to measure the validity of the item, with the test as a whole as criterion, he will be assuming that the distribution of scores on the test is very like the normal curve and that the item is actually measuring a trait which is continuous, normally distributed, and has rectilinear regression on the test. If these assumptions are not all fulfilled, he may in fact be rejecting some of the better and keeping some of the worse items.

Lastly, in normalizing the distribution of scores on the final form of the test, he may believe he is infusing some mysterious property of "additivity" or "equal units of measurement" into the test. But we hope he will claim no more than that the normalized scale is independent of the original test and of the sample used in normalizing. His claim that normalized scores are independent of the sample has not yet been convincingly demonstrated, but such a demonstration may be possible. More important is that his whole attempt at scaling the scores as-

sumes that this test is homogeneous in the ability or abilities tested, and he has offered us no information on the homogeneity of the test except the ambiguous evidence from reliability coefficients.

In contrast with the hypothetical Psychologist Smith, let us consider what is being done by an even more hypothetical Psychologist Jones, who constructs a test in accordance with the ideas proposed in the second part of this essay. He will evaluate the original form of the test in terms of an index of its homogeneity. He will seek to improve the test by eliminating the items which are least homogeneous with the whole test. Only after he has largely succeeded in producing a homogeneous test will he attempt to scale the scores; then he will insist that the method of scaling be such that the resultant scores are independent of the original test and of the original sample.

Throughout, Psychologist Jones will assume that the ability or abilities measured by his test can be ascertained at various levels of difficulty, that is, that changing the difficulty of the item does not necessarily change the nature of the ability measured, and he will assume that the abilities of the various items of his test are not negatively related to each other, that is, that the abilities are related either positively or not at all. The first assumption is the basis for the attempt to construct homogeneous tests; if it is wrong, the psychologist will not be led into a wrong interpretation of facts but will fail in achieving his goal just to the extent that the assumption is wrong. The second assumption is used as a basis for defining a perfectly heterogeneous test, the result of complete failure to achieve the goal set; if this assumption is wrong, the reference point of perfect heterogeneity will be wrong or meaningless. But this assumption is in accordance with a large amount of experience obtained in several decades of testing abilities.

The notion of homogeneity is based not on a division of scores into unseen "true scores" and "error factors" but purely on the answers to the items. A test is perfectly homogeneous if answering one item correctly implies answering all previous items correctly, when items are arranged in order of difficulty. A test is perfectly heterogeneous if there is no relation between answering one item correctly and answering other items correctly. The notion of "chance" enters not as a description of the mental processes of the person taking the test, but as a description of the relation between answers to the items of a hypothetical perfectly heterogeneous test. There need be no such perfectly heterogeneous test. It is used only as a reference point in evaluating how far the given test has achieved homogeneity.

While no doubt various indices of homogeneity could be devised, the one that has been recommended has the virtue of simplicity in computation and in conception. The computational form is based on a demonstration that the variance of a test is a function only of the difficulties of the items and of the proposed index of homogeneity. The variance of the test thus is shown to be an index of its homogeneity, when expressed on a scale with the variance of a perfectly heterogeneous test with the same item difficulties as lower limit, and the variance of a perfectly homogeneous test with the same item difficulties as upper limit.

The measurement of the homogeneity of the item with the test involves essentially the same assumptions and the same concepts as the measurement of the homogeneity of the test. No restriction

is placed on the form of distribution of the test, and no assumption is made about the "trait" measured by the item; there are just right responses and wrong responses. In a perfectly homogeneous test, all individuals with right responses to the item will have higher total scores than all individuals with wrong responses to the item. In a perfectly heterogeneous test, an individual who passes the item will be as likely to be lower in total score as higher than an individual who fails the item. The index of homogeneity of item with test is equivalent to the number of right discriminations minus the number of wrong discriminations divided by the total number of discriminations, where every pair of persons who differ in item score and in total score is one "discrimination". This index equals one for an item which is perfectly homogeneous with its test, and is never greater than one. It was not shown equal to zero in the case of an item perfectly heterogeneous with its test; an adequate index would also have this property.

Psychologist Jones so far has made use of the rank order of scores rather than their magnitude, while Psychologist Smith utilized the actual magnitude of the scores in reporting reliability and in measuring item validity. Before scaling his test, Psychologist Jones will demand reasonable fulfillment of the criterion of perfect homogeneity. Exactly how far a test can be made homogeneous, and how far the obdurate vagaries of human performance will prevent achieving that goal, experience must answer. Since the index of homogeneity is computed from a single administration of a test, however, the effect of variations within the individual certainly is minimized.

An acceptable method of scaling must result in a derived scale which is independent of the original scale and of the original group tested, "independence" being interpreted in the sense of perfect correlation with scales derived from different tests of the same ability or from groups differing in mean performance. In lieu of analytical proof that a given method of scaling is adequate, empirical evidence may have to be accepted. Such evidence must be based on reasonably homogeneous tests, and it will be required that the derived scales not be highly correlated with the raw scores, in order that the high correlation between scales derived from different groups be certainly not accidental.

Until an adequate system of scaling is found, the correlation between tests of abilities, even between two tests of the same ability, will be accidental to an unknown degree. Since the method of factor analysis depends on the correlation between tests to discover the relationships between abilities, further refinement in the precision of the results of factor analyses appears to depend on the solution of the problem of scaling. If, ultimately, no solution to the problem of scaling is found, alternatives to the traditional correlational analyses must be sought.

In defense of Psychologist Smith, it should be noted that nothing has been said in this thesis about the use of tests of ability to predict success in life situations, nor does the criticism of the use of correlation apply to this instance. Tests constructed by his methods have a long history of value for a variety of purposes, and it remains to be seen how tests constructed according to the proposed systematic test statistics will compare in final validity.

## Appendix

### Derivation of the computational form of $H_t$

Consider first the formula for the variance of a perfectly homogeneous test, formula (9):

$$V_{hom} = \sum_{i=1}^{m} p_i q_i + 2\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (p_j - p_i p_j).$$

Substituting for $q_i$ its value, $1 - p_i$, and noting that the sum of the differences is equal to the difference of the sums, we obtain:

$$V_{hom} = \sum p_i - \sum p_i^2 + 2\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} p_j - 2\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} p_i p_j.$$

Where the limits of summation are not indicated, it will be understood the summation extends for the $m$ items. The second and fourth terms in the last expression are simply the expansion of $(\Sigma p_i)^2$. If twice the first term is added to the third term, we obtain:

$$V_{hom} = 2\sum_{i=1}^{m} \sum_{j=i}^{m} p_j + \sum p_i - 2\sum p_i - (\sum p_i)^2.$$

Expansion of the first term yields $2\Sigma i p_i$. Therefore,

$$V_{hom} = 2\sum i p_i - \sum p_i - (\sum p_i)^2.$$

Substituting this value for $V_{hom}$ and the value obtained in formula (10) for $V_{het}$ into formula (7) for $H_t$, we have:

$$H_t = \frac{V_x - V_{het}}{V_{hom} - V_{het}} = \frac{V_x - \sum p_i q_i}{2\sum i p_i - \sum p_i - (\sum p_i)^2 - \sum p_i q_i}$$

$$= \frac{V_x - \sum p_i + \sum p_i^2}{2\sum i p_i - \sum p_i - (\sum p_i)^2 - \sum p_i + \sum p_i^2}$$

$$= \frac{V_x - \sum p_i + \sum p_i^2}{2(\sum i p_i - \sum p_i) - (\sum p_i)^2 + \sum p_i^2}.$$

As suggested in the text, for large samples it will probably prove to be satisfactory to substitute the variance of the sample, $V_{sample}$, for the variance of the population and to substitute the proportion of the sample passing the item, $P_i$, for the probability of passing the item, $p_i$. The order of difficulty, denoted by $i$, must also be estimated from the sample. This procedure will yield an estimate of the homogeneity which we may denote by *Est* $H_t$.

$$\text{Est } H_t = \frac{V_{sample} - \sum P_i + \sum P_i^2}{2(\sum i P_i - \sum P_i) - (\sum P_i)^2 + \sum P_i^2}$$

For computational purposes it is simpler to deal with numbers than with proportions. Denote by $N_i$ the number passing the $i$th item. Obviously $N_i$ is equal to $Np_i$. Substituting these values, multiplying numerator and denominator by $N^2$,

and substituting the value of $V_{sample}$ in terms of scores, we have:

$$\text{Est } H_i = \frac{N\sum X_k^2 - (\sum X_k)^2 - N\sum N_i + \sum N_i^2}{2N(\sum iN_i - \sum N_i) - (\sum N_i)^2 + \sum N_i^2},$$

where the subscript $k$ denotes summation for the $N$ individuals. But the total number of correct responses is the same whether summed first by persons or first by items. Thus $\Sigma N_i$ must exactly equal $\Sigma X_k$. Making this substitution and rearranging terms yields:

$$\text{Est } H_i = \frac{N(\sum X_k^2 - \sum X_k) + \sum N_i^2 - (\sum X_k)^2}{2N(\sum iN_i - \sum X_k) + \sum N_i^2 - (\sum X_k)^2}.$$

## References


1. Bergmann, Gustav, and Spence, Kenneth W. The logic psychophysical measurement. *Psychol. Rev.*, 1944, *51*, 1-24.
2. Brown, William, and Thomson, Godfrey H. The essentials of mental measurement. (4th ed.) Cambridge: Cambridge Univ. Press, 1940.
3. Buros, Oscar K. (Ed.) The Nineteen Forty Mental Measurement Yearbook. Highland Park, N.J.: Gryphon Press, 1940.
4. Cattell, Raymond B. The measurement of adult intelligence. *Psychol. Bull.*, 1943, *40*, 153-193.
5. Dressel, Paul L. Some remarks on the Kuder-Richardson reliability coefficient. *Psychometrika*, 1940, *5*, 305-310.
6. Flanagan, John C. The Cooperative Achievement tests: A bulletin reporting the basic principles and procedures used in the development of their system of scaled scores. New York: Cooperative Test Service, 1939.
7. Flanagan, John C. General considerations in the selection of test items and a short method of estimating the product-moment coefficient from data at the tails of the distribution. *J. educ. Psychol.*, 1939, *30*, 674-680.
8. Flanagan, John C. Statistical methods related to test construction and evaluation. *Rev. educ. Res.*, 1941, *11*, 109-130.
9. Froehlich, Gustav J. A simple index of test reliability. *J. educ. Psychol.*, 1941, *32*, 381-385.
10. Goodenough, Florence L. A critical note on the use of the term "reliability" in mental measurement. *J. educ. Psychol.*, 1936, *27*, 173-178.
11. Greene, Edward B. Measurements of human behavior. New York: Odyssey Press, 1941.
12. Grossnickle, Louise T. The scaling of test scores by the method of paired comparisons. *Psychometrika*, 1942, *7*, 43-64.
13. Guilford, J. P. Human abilities. *Psychol. Rev.*, 1940, *47*, 367-394.
14. Gulliksen, Harold. A course in the theory of mental tests. *Psychometrika*, 1943, *8*, 223-245.
15. Horst, Paul. Item selection by means of a maximizing function. *Psychometrika*, 1936, *1*, 229-244.
16. Hoyt, Cyril. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, *6*, 153-160.
17. Jackson, Robert W. B., and Ferguson, George A. Studies on the reliability of tests. Bull. No. 12, Dept. Educ. Res., University of Toronto, 1941.
18. Kelley, Truman L. The reliability coefficient. *Psychometrika*, 1942, *7*, 75-83.
19. Kendall, M. G. A new measure of rank correlation. *Biometrika*, 1938, *30*, 81-93.
20. Kuder, G. F., and Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, *2*, 151-160.
21. Long, John A. Improved overlapping methods for determining the validities of test items. *J. exper. Educ.*, 1934, *2*, 264-268.
22. Long, John A., Sandiford, Peter, and others. The validation of test items. Bull. No. 3, Dept. Educ. Res., University of Toronto, 1935.
23. Pearson, Karl. On the mathematical theory of errors of judgment, with special reference to the personal equation. *Philos. Trans.*, A, 1902, *198*, 235-299.
24. Richardson, M. W. The relation between the difficulty and the differential validity of a test. *Psychometrika*, 1936, *1*, 33-49.
25. Richardson, M. W. Notes on the rationale of item analysis. *Psychometrika*, 1936, *1*, 69-76.
26. Richardson, M. W., and Stalnaker, J. N. A note on the use of biserial r in test research. *J. gen. Psychol.*, 1933, *8*, 463-465.
27. Rosander, A. C. The use of inversions as a test of random order. *J. Amer. statist. Ass.*, 1942, *37*, 352-358.
28. Spearman, C. Correlation calculated from faulty data. *Brit. J. Psychol.*, 1910, *3*, 271-295.
29. Symonds, Percival M. Choice of items for a test on the basis of difficulty. *J. educ. Psychol.*, 1929, *20*, 481-493.
30. Thomson, Godfrey H. The nature and measurement of intellect. *Teach. Coll. Rec.*, 1940, *41*, 726-750.
31. Thorndike, Edward L., *et al.* The measurement of intelligence. New York: Teachers College, Columbia Univ., 1926.
32. Thurstone, L. L. A method of scaling psychological and educational tests. *J. educ. Psychol.*, 1925, *16*, 433-451.
33. Thurstone, L. L. The unit of measurement in educational scales. *J. educ. Psychol.*, 1927, *18*, 505-524.
34. Thurstone, L. L. The reliability and validity of tests. Ann Arbor: Edwards, 1931.
35. Thurstone, L. L. Current misuse of the factorial methods. *Psychometrika*, 1937, *2*, 73-76.
36. Thurstone, Thelma Gwinn. The difficulty of a test and its diagnostic value. *J. educ. Psychol.*, 1932, *23*, 335-343.
37. Wherry, Robert J., and Gaylord, Richard H. The concept of test and item reliability in relation to factor pattern. *Psychometrika*, 1943, *8*, 247-264.